Blame it on me

Lambèr Royakkers¹ · Jesse Hughes²

Received: 9 September 2016 / Accepted: 24 May 2019 / Published online: 10 June 2019
© The Author(s) 2019

Abstract

In this paper, we develop a formalisation of the main ideas of the work of Van de Poel on responsibility. Using the basic concepts through which the meanings of responsibility are defined, we construct a logic which enables to express sentences like “individual i is accountable for φ ”, “individual i is blameworthy for φ ” and “individual i has the obligation to see to it that φ ”. This formalization clarifies the definitions of responsibility given by Van de Poel and highlights their differences and similarities. It also helps to assess the consistency of the formalisation of responsibility, not only by showing that definitions are not inconsistent, but also by providing a formal demonstration of the relation between three main meanings of responsibility (accountability, blameworthiness, and obligation). The formal account can be used to derive new properties of the concepts. With the help of the formalisation, we detect the occurrence of the problem of many hands (PMH) by defining a logical framework for reasoning about collective and individual responsibility. This logic extends the Coalition Epistemic Dynamic Logic (CEDL) by adding a notion of group knowledge (and generalize the definitions of individual responsibility to groups of agents), agent ability and knowing how to its semantics.

Keywords Responsibility · Coalition epistemic logic · Problem of many hands · Collective agency

✉ Lambèr Royakkers
l.m.m.royakkers@tue.nl

Jesse Hughes
jesse@phiwumbda.org

¹ Faculty of Industrial Engineering and Innovation Sciences, Eindhoven University of Technology, Eindhoven, The Netherlands

² Philosophy Department, Bentley University, Waltham, MA, USA

1 Introduction

The concept of responsibility is central to a theory of collective agency and organisations. Responsibility issues arise any time a group of agents acts collectively in order to achieve certain objectives. Plans are made for the collective action of the group and specific agents are stated to be “responsible” for certain tasks. If something goes wrong certain agents might be found “responsible” for what happened, they might be held “accountable” and be “blamed”. The notions of responsibility display different nuances all related with particular aspects of, such as, agents’ ability, obligation and knowledge.

Van de Poel [38, 39] has presented a conceptual analysis of moral responsibility. He has introduced a taxonomy of the various normative meanings of responsibility. For example, when we say that someone is responsible for a consequence φ , it may mean that one is “accountable” for φ , or that one is “blameworthy” for φ , or, that one has the “obligation to see to it that” φ . As we have also seen the application of these various meanings of responsibility depends on more basic concepts, such as moral agency, causality, wrong-doing, freedom and knowledge. For instance, an individual i is accountable for φ if i is capable of acting as a moral agent, is obliged to realise $\neg\varphi$ but instead acts so as to necessitate φ . In addition, i is blameworthy for φ if i is accountable for φ , knows (or could know) that φ would be the case, and acts freely, i.e., i could have chosen a morally acceptable plan which would not have necessitated φ . Finally, i has the obligation to see to it that φ if i must ensure the occurrence of φ .

Van de Poel has classified these different meanings of responsibility in two categories. Accountability and blameworthiness are “backward-looking responsibilities”, i.e., responsibilities regarding the past, whereas “obligation to see to it that” is a “forward-looking responsibility”, i.e., a responsibility regarding the future. Van de Poel [39] also argues that these three notions are related, and has given four conceptual relations between these different notions. Roughly, (1) if an individual i has the obligation to see to it that φ but does not realise φ , then i is accountable for $\neg\varphi$; (2) if i is accountable for φ and does not provide a satisfactory account for φ , then i is blameworthy for φ , (3) if i is blameworthy for φ , then i is also accountable for φ , and (4) if φ is not the case and i is not accountable for $\neg\varphi$, then i had no obligation to see to it that φ .

Using the basic concepts upon which the meanings of responsibility are defined, we construct a logic which enables to express sentences like “individual i is accountable for φ ”, “individual i is blameworthy for φ ” and “individual i has the obligation to see to it that φ ”. Such effort contributes to the discussion about responsibility in at least two ways. First, it clarifies the definitions and also their differences and similarities. Second, it assesses the consistency of formalisation of responsibility, not only by showing that definitions are not inconsistent, but also by providing a formalisation of the four relations between the three notions of responsibility. Moreover, the formal account can be used to derive new properties of the concepts, thus, giving new insights that can be used to advance the discussion. And finally, the formalism proposed here provides a framework wherein criteria for ascribing responsibilities can be stated and, if individuals are to be held responsible for outcomes, then, at least, justifications can be made clear.

We also develop a formal theory in which the “problem of many hands” (PMH), a term coined by Thompson in [37], can be precisely understood. This problem has been studied in the context of organisations (e.g., by Bovens [8]) and has to do with the potential discrepancy between individual and collective responsibility: in some situations, a group of individuals can reasonably be held responsible for an undesirable outcome, while no member of the group can reasonably be held responsible for the outcome. Bovens argues that such situations are problematic because it “frustrates the need for compensation and retribution on the part of the victims” and, maybe more important, because “the fact that no one can be meaningfully called to account after the event also means . . . that no one need feel responsible beforehand” [8]. In our formalisation, we are able to precisely define when a group is responsible (both in terms of accountability and blameworthiness) and also when individuals (or subgroups, more generally) share that responsibility.

In this paper, we define a logical framework for reasoning about collective and individual responsibility based on Coalition Epistemic Dynamic Logic (CEDL) proposed in [13]. However, the formalisation found in de Lima, et al., was not completely adequate, because it misses, for example, *indirect accountability* which is, especially in organisations, an essential notion with respect to responsibility. We show this here through the example “Light bulb and light switch”. Consider a scenario inhabited by two agents: Alice and Betty. They live in a strange house: its interior is illuminated by a light bulb, but the corresponding switch is located outside the house. In this scenario, Alice is inside the house and Betty is outside it, close to the switch. Thus, Alice can see whether the light bulb is on or off and tell (or rather shout) this fact to Betty, but she cannot toggle the switch. Betty, on the other hand, can toggle the switch but she cannot see whether the light is on or off. If she toggles the switch with the light on, it will turn off, and if she toggles it with the light off, it will turn on. Suppose that the light bulb is off and that Alice and Betty have the obligation that the light be on. If Alice and Betty do nothing, they are in a violation state. Betty is accountable, *directly accountable*, since she could perform an action (toggle the switch) that necessarily leads to the fact that the light bulb is on. On the other hand, Alice could have informed Betty that the light bulb is off. In this case Betty would have the information to prevent a violation. So, one may conclude that if both do nothing, ending in a violation state, it is Alice’s fault as well, because she is *indirectly accountable*. This intuitive conclusion cannot be achieved using the formalisation proposed in [13]. Furthermore, we say that both individuals are accountable, but that only Alice is blameworthy, since she knew how the group (Alice and Betty) could have prevented the violation. In this paper we will address these issues raised by this example, among others.

To provide a formal understanding of responsibility issues within group of agents is of definite importance for the theory and development of multi-agent systems (MAS). In fact, many methodologies for MAS are based on organisational concepts as their cornerstones. A formal theory of responsibility would provide these methodologies with conceptual tools for interpreting, in organisational terms, faulty performances of a given MAS, and at the same time suggest guidelines for the design of MAS behaving in specific ways with respect to the assessment of responsibilities among the agents.

The rest of this paper is organised as follows. In the next section we will introduce a variation of CEDL. Then, in Section 3, we define the operator to express agent ability and obligation. Sections 4, 5 and 6, present the formalisation of three notions of responsibility, namely, obligation to see to it that, accountability, and blameworthiness. In Section 7, we will provide the formalisation of Van de Poel's relations between the three notions of responsibility. The last section draws conclusions and discusses possible future work.

2 The Formal Framework

In this section we present a logic that will be the basis of our formalisation of individual responsibility. This logic is a variation of the coalition epistemic dynamic logic (CEDL),

which, in turn, is an extension of the well-known propositional dynamic logic (PDL) [21]. PDL is a classical propositional logic augmented with modal operators ' $[a]$ '. A formula of the form $[a]\varphi$ means 'after every possible occurrence of event a , the consequence is true'. Thus, it permits expression of what consequences are caused by the occurrence of some given event a , where such events can be actions executed by one agent, actions executed by several agents, exogenous events or even programs. But since PDL does not have agents in its language, it does not enable expression of what consequences are caused by which agents of the scenario. To be able to express agent causality, CEDL extends it by actions that are 'enacted' by agents, in a similar way as done, e.g., Wieringa and Meyer [41], and more recently by Lorini and Herzig [22, 26] and Ågotnes and Alechina [2]. In CEDL, one can write formulae of the form $[(i, a)]\varphi$, meaning 'after every possible occurrence of event (i, a) , the consequence is true', where the event (i, a) is the action a executed by agent i . Moreover, to be able to express agent knowledge, CEDL has modal operators K_i , in a similar way as done, e.g., by Grossi et al. [20] and Herzig et al. [23]. In CEDL, one can also write formulae of the form $K_i\varphi$, meaning 'agent i knows that φ '.

The remainder of this section is organised as follows. The next subsection presents the models used to give semantics to CEDL. Then, Section 2.2 presents CEDL language, its interpretation on the models and its complete axiomatisation. After that, Section 2.3 provides a formalisation of group knowledge. And Finally, Section 2.4 compares the models of CEDL with the Concurrent Epistemic Game Models (CEGMs).

2.1 Models

The mathematical structures defined in this section aim at modeling environments inhabited by one or more entities called agents. These agents are able to execute actions in order to change the environment. These include epistemic actions, i.e., actions that change the knowledge of the agents of the environment, such as sensing actions, communication actions, etc., and also physical actions, i.e., actions that change the environment, such as toggling light switches, opening and closing doors, etc. The idea is to be able to describe scenarios where one may want to ascribe

responsibilities to agents or simply want to verify whether agents are responsible or not for consequences. Then, ideally, these structures should be rich enough to express the important concepts that define the various conditions of responsibility discussed in [39], namely, moral agency, causality, wrong-doing, freedom and knowledge. However, as the name suggests, these structures are meant to “model” the reality and not to copy it. Hence, some simplifications are made.

A model is defined for a given vocabulary, which consists of a triple of disjoint set $\langle P, N, A \rangle$ where:

- P is a countable (possibly infinite) set of propositional variables denoting propositional facts;
- N is a finite set of labels denoting agents; and
- A is a finite set of labels denoting the actions available for the agents.

We use Δ to denote the set of all joint actions available for the agents in N , which is defined as the set of all total functions δ with signature $N \rightarrow A$. In other words, $\Delta = \{\delta_1, \delta_2, \dots\}$, where each δ_j is a set of pairs of the form $\{(i_1, a_1), (i_2, a_2), \dots, (i_{|N|}, a_{|N|})\}$ (one pair for each agent in N) and where $i_m \in N$ and $a_m \in A$ for $1 < m < |N|$.

CEDL models are a specific kind of Kripke models, which consist of directed labelled graphs as structures, where nodes represent possible worlds and relations between worlds represent (in our case) either indistinguishability relations or transitions. The indistinguishability relation for agent i represents the knowledge of i : the more worlds are indistinguishable for i , the less agent i knows about the scenario. The transition relation for some action δ represents the set of possible outcomes of δ : if w_1 is in the transition relation of δ with w_0 then it means that the occurrence of δ at w_0 may lead to w_1 . A graphical representation of one such model is displayed in Fig. 1.

In such a structure, each possible world is labelled with a set of propositional variables from P , each indistinguishability relation is labelled with an agent from N , and each transition is labelled with a joint action from N . For example, consider the

Fig. 1 A graphic representation of a CEDL model. (Circles represent possible worlds, arrows represent transitions and rectangles enclose worlds that belong to the same indistinguishability relation)

model in Fig. 1. The propositional variable p is present in the label of w_0 and w_2 but it is not present in the label of w_1 . This means that the propositional fact represented by p is true at w_0 and w_2 , but it is false at w_1 . Moreover, the indistinguishability relation between w_0 and w_1 is labelled with agent i , meaning that agent i cannot distinguish between these two worlds. Thus in our example, if w_0 (or w_1) is the actual world, then i considers it possible that either w_0 or w_1 is the actual world. This means that, if w_0 (or w_1) is the actual world, then i does not know whether the fact represented by p is true or false. And finally, the transition between w_0 and w_2 is labelled with joint action δ , meaning that the occurrence of δ at w_0 leads to w_2 .

The formal definition is as follows. Let the vocabulary $\langle P, N, A \rangle$ be given, a CEDL model is a quadruple $\langle W, \mathcal{K}, \mathcal{T}, \mathcal{V} \rangle$, where:

- W is a non-empty set of possible worlds;
- $\mathcal{K}$ is a function with signature $N \rightarrow \mathcal{P}(W \times W)$, defining, for each agent $i \in N$, an equivalence relation between possible worlds, which represents the knowledge of agent i ;
- $\mathcal{T}$ is a function with signature $\Delta \rightarrow \mathcal{P}(W \times W)$, defining, for each joint action $\delta \in \Delta$, a relation between possible worlds, which represents the transition associated to the joint action δ ; and
- $\mathcal{V}$ is a function with signature $P \rightarrow \mathcal{P}(W)$, defining, for each $p \in P$, the interpretation of the propositional variable p in the model.

To simplify notation, we write $\mathcal{K}_i$ instead of $\mathcal{K}(i)$, and also use $\mathcal{K}_i(w)$ to denote the set of worlds that i considers possible at w . That is, $\mathcal{K}_i(w)$ abbreviates the set $\{w' : (w, w') \in \mathcal{K}(i)\}$. Analogously, we write $\mathcal{T}_\delta$ instead of $\mathcal{T}(\delta)$, and also use $\mathcal{T}_\delta(w)$ to denote the set of possible outcomes of the occurrence of δ at w . That is, $\mathcal{T}_\delta(w)$ abbreviates the set $\{w' : (w, w') \in \mathcal{T}(\delta)\}$.

We impose the following constraint to CEDL models:

$$\text{for all } w \in W \text{ there is } \delta \in \Delta \text{ such that } \mathcal{T}_\delta(w) \neq \emptyset \quad (\text{C1})$$

Constraint (C1) is called ‘activity’. It stipulates that, at all possible worlds of the model, there is at least one non-empty transition which is labelled by some joint action in Δ . In other words, at any moment, there is at least one executable action for each agent, which prevents those strange structures mentioned earlier from being CEDL models.

The dynamic portion of our model is quite similar to the joint action models over Π_{ActN}^0 discussed in [2].¹ There are two primary differences to note.² First, joint action models are explicitly deterministic, a requirement we omit. We allow that the outcome of the whole group N may be non-deterministic, since agents may perform actions with unknown outcome (one member of the group may throw a ball at a target, for instance, with no guarantee that it hits the bullseye). Determinism, in our setting, amounts to the requirement that $\mathcal{T}_\delta(w)$ is a singleton for every δ and w .

¹The addition of epistemic operators to joint action models would be straightforward, along the lines discussed in [1].

²An additional minor difference is that joint action models include inconsistent actions, in which an agent performs two distinct actions simultaneously. This is a trivial difference, since such inconsistent actions are executable nowhere.

Secondly, joint action models satisfy the following (axiom **A3** in *ibid*, also referred to as *Independent Choice* (IC)): if, in world w , Stan can perform a given action a_S and Ollie, in the same world, can perform a_O , then it is possible for Stan to perform a_S and Ollie a_O simultaneously. However, Stan may be able to walk through a doorway if and only if Ollie doesn't walk through the doorway at the same time (or else they get stuck and fail to perform the action "walk through the doorway"). Similarly, Will may write a play with a pencil or Abe may write an address with the same pencil, but they can't do these two actions simultaneously. We will discuss IC in some detail in Section 5.1 and in particular indicate why we find it unsuitable for our aims.

There is yet another constraint imposed to CEDL models:

$$\text{for all } w \in W, \text{ all } i \in N \text{ and all } \delta \in \Delta, \text{ we have } (\mathcal{T}_\delta \circ \mathcal{K}_i)(w) \subseteq (\mathcal{K}_i \circ \mathcal{T}_\delta)(w) \quad (\text{C2})$$

This constraint corresponds to 'no-forgetting' in [23] and 'perfect recall' in [17]. It defines an interaction between accessibility relations and transitions. With this constraint, the knowledge of an agent either increases or stays the same, after the execution of any action. This means that agents never lose information, i.e., once an agent knows something, this agent will never forget it. This is obviously very restrictive but, at the same time, useful. For instance, this assumption allow us to ignore models where agents may keep losing information for whatever reason. Such possibility would make much more difficult (if not impossible) to derive some interesting properties of our formalism. We also note that, because each $\mathcal{K}_i$ is an equivalence relation, constraint (C2) implies that action occurrences are perceived by all agents. The latter implies that the agents perceive the passage of time.

For example, the structure in Fig. 1 respects both (C1) and (C2). The first is respected because, for all possible worlds w , there is at least one transition from w to some possible world. The second is respected because the equivalence classes of the accessibility relation $\mathcal{K}_i$ does not increase its size when we follow a transition from one possible world to another.

2.2 Syntax and Semantics of CEDL

Given a vocabulary $\langle P, N, A \rangle$, the language of CEDL is the smallest set $\mathcal{L}$ which satisfies the following conditions:

- $\top \in \mathcal{L}$;
- if $p \in P$ then $p \in \mathcal{L}$;
- if $\varphi \in \mathcal{L}$ then $\neg\varphi \in \mathcal{L}$;
- if $\varphi \in \mathcal{L}$ and $\psi \in \mathcal{L}$ then $\varphi \wedge \psi \in \mathcal{L}$;
- if $i \in N$ and $\varphi \in \mathcal{L}$ then $K_i\varphi \in \mathcal{L}$;
- if $\delta \in \Delta$, $G \subseteq N$ and $\varphi \in \mathcal{L}$ then $[\delta|_G]\varphi \in \mathcal{L}$,

where $\delta|_G$ is δ , but with its domain restricted to G , i.e., let δ be the joint action $\{(i_1, a_1), \dots, (i_N, a_N)\}$, then $\delta|_G$ is the partial joint action formed by the set of pairs $\{(i_n, a_n) \mid (i_n, a_n) \in \delta \text{ and } i_n \in G\}$.

In what follows, the common abbreviations for the operators $\vee$, $\rightarrow$ and $\leftrightarrow$ are also used, the symbol $\perp$ (contradiction) abbreviates $\neg\top$, and, to simplify notation,

we sometimes write $[i_1:a_1, \dots, i_n:a_n]\varphi$ instead of $\{[(i_1, a_1), \dots, (i_n, a_n)]\}\varphi$. We also sometimes write $[\delta_1|_G; \dots; \delta_n|_G]$ to mean $[\delta_1|_G] \dots [\delta_n|_G]$.

Note also that, as Δ , N and 2^N are finite sets, existential and universal quantification over these sets abbreviate disjunctions and conjunctions, respectively.

The intended meaning of formulae of the form $K_i\varphi$ is, as usual: ‘agent i knows that φ ’. The meaning of a partial joint action of the form $\{(i_1, a_1), \dots, (i_n, a_n)\}$ is: ‘the agents in $\{i_1, \dots, i_n\}$ execute their corresponding actions in $\{a_1, \dots, a_n\}$ simultaneously (and we do not consider what the other agents do at the same time)’. And the meaning of formulae of the form $[\delta|_G]\varphi$ is: ‘after every possible occurrence of δ_G , φ is true’. Furthermore, $\langle\delta|_G\rangle\varphi$ abbreviates $\neg[\delta|_G]\neg\varphi$.

To match their meanings, formulae from $\mathcal{L}$ are interpreted using ‘pointed CEDL models’. The latter are pairs of the form (M, w) , where $M = \langle W, \mathcal{K}, \mathcal{T}, \mathcal{V} \rangle$ is a CEDL model and $w \in W$. The semantic interpretation of Boolean operators is the usual one. For formulae of the form $K_i\varphi$ we use the accessibility relation $\mathcal{K}_i$: $K_i\varphi$ is true at the world w if and only if φ is true at all possible worlds w' accessible from w via $\mathcal{K}_i$, i.e., it is true if and only if φ is true at all worlds that the agent i considers possible at w . The interpretation of operators $[\delta|_G]$ is more complex. A formula of the form $[\delta|_G]\varphi$ is true at the world w if and only if φ is true at all possible worlds w' that are attained from w via some transition in $\mathcal{T}$ which is labelled with $\delta|_G \cup \delta'|_{N \setminus G}$. In other words, to verify whether $[\delta|_G]\varphi$ is true at some possible world w , we must verify whether φ is true at all worlds w' belonging to $\mathcal{T}_{\delta'}(w)$ for all $\delta' \in \Delta$, provided that the restriction of δ' to G is equal to the restriction of δ to G , i.e., provided that $\delta'|_G = \delta|_G$. That is, we must verify whether φ is true at all G transitions from w no matter what the agents outside the group G do.

We say that an action δ *extends* $\delta'|_G$ (written $\delta \succcurlyeq \delta'|_G$) just in case $\delta|_G = \delta'|_G$. Clearly, if $\delta \succcurlyeq \delta'|_G$ and $G' \subseteq G$, then $\delta \succcurlyeq \delta'|_{G'}$.

Similarly, we say that a sequence $\delta_1; \dots; \delta_n$ *extends* $\delta'_1|_G; \dots; \delta'_n|_G$ (written $\delta_1; \dots; \delta_n \succcurlyeq \delta'_1|_G; \dots; \delta'_n|_G$) just in case each $\delta_i|_G$ extends $\delta'_i|_G$.

Formally, the satisfaction relation $\models$, between pointed CEDL models and formulae from $\mathcal{L}$, is inductively defined as follows:

$$\begin{aligned}
 M, w &\models \top \\
 M, w &\models p && \text{iff } w \in \mathcal{V}(p) \\
 M, w &\models \neg\varphi && \text{iff } M, w \not\models \varphi \\
 M, w &\models \varphi \wedge \psi && \text{iff } M, w \models \varphi \text{ and } M, w \models \psi \\
 M, w &\models K_i\varphi && \text{iff for all } w' \in \mathcal{K}_i(w) \text{ we have } M, w' \models \varphi \\
 M, w &\models [\delta|_G]\varphi && \text{iff for all } \delta' \succcurlyeq \delta|_G \text{ and all } w' \in \mathcal{T}_{\delta'}(w) \text{ we have } M, w' \models \varphi
 \end{aligned}$$

As usual, a formula φ is valid (notation: $M \models \varphi$ or, when context is clear, $\models \varphi$) if and only if every pointed CEDL model (M, w) satisfies φ .

Example 1 (Light Bulb and Light Switch) To better explain the definitions given so far, we now use the ‘light bulb and light switch’ scenario from the Introduction. Alice is inside the house and Betty is outside it, close to the switch. Thus, Alice can see whether the light bulb is on (denoted p) or off (denoted $\neg p$) and tell it to Betty (action *tell*), but she cannot toggle the switch. Betty, on the other hand, can

toggle the switch (action *tog*) but she cannot see whether the light is on or off. If she toggles the switch with the light on, it will turn off, and if she toggles it with the light off, it will turn on. Let the vocabulary be $\langle P, N, A \rangle$, where $P = \{p\}$, $N = \{a, b\}$ and $A = \{none, tell, tog\}$. A CEDL model for this scenario is given by the structure in Fig. 2.

We assume that the light bulb is off, i.e., the actual world is w_0 . We then use the definition of $\models$ to verify truth of some formulae in the pointed CEDL model (M, w_0) :

- $M, w_0 \models \neg p$. In words, the light is off. This is true because $w_0 \notin \mathcal{V}(p)$, i.e., p is not in the label of w_0 .
- $M, w_0 \models \neg K_b p \wedge \neg K_b \neg p$. Betty does not know whether the light is on or off. This is true because there is a world that Betty considers possible (namely w_0) where the light is off and there is a possible world that Betty considers possible (namely w_1) where the light is on.

Many other interesting (and more complex) formulae can be verified in this model. We leave it to the reader to verify the following formulae:

- $M, w_0 \models [a: tell, b: tog] p$: after the parallel execution of Alice telling whether the light is on or off and Betty toggling the switch, the light is on.
- $M, w_0 \models [b: tog] p$. After Betty toggles the switch, the light is on. (Note that here we consider only the action executed by Betty.)
- $M, w_0 \models \neg[a: tell] K_b \neg p$. It is not the case that after Alice tells Betty that the light is off, she knows that the light is off. (This is because Betty might have toggled the switch at the same time, so the light might actually be on.)
- $M, w_0 \models [a: tell, b: none] K_b [b: tog] (K_a p \wedge K_b p)$. After the parallel execution of Alice telling whether the light bulb is on or off to Betty and Betty doing nothing, Betty knows that after Betty toggling the switch, the two agents know that the light is on. In other words, it means that if Betty waits for the announcement of Alice, then she knows how to reach the state where the two agents know that the light is on.

Fig. 2 A CEDL model for the Light Bulb and Light Switch scenario. (The rectangles represent Betty's knowledge. To simplify the picture, Alice's knowledge is not represented. She has complete knowledge of the scenario, i.e., $\mathcal{K}_a(w) = \{w\}$ for all $w \in W$)

Table 1 The axiomatisation of CEDL

All instances of the propositional tautologies		(TAU)
$(K_i \varphi \wedge K_i (\varphi \rightarrow \psi)) \rightarrow K_i \psi$	(deductive closure for knowledge)	(KK)
$K_i \varphi \rightarrow \varphi$	(truth)	(T)
$K_i \varphi \rightarrow K_i K_i \varphi$	(positive introspection)	(4)
$\neg K_i \varphi \rightarrow K_i \neg K_i \varphi$	(negative introspection)	(5)
$([\delta _G] \varphi \wedge [\delta _G] (\varphi \rightarrow \psi)) \rightarrow [\delta _G] \psi$	(deductive closure for action)	(KA)
$\exists \delta \in \Delta \neg [\delta _G] \perp$	(activity)	(A)
$\forall \delta' \succcurlyeq_{\delta _G} [\delta'] \varphi \rightarrow [\delta _G] \varphi$	(deriving action)	(DA)
$K_i [\delta _G] \varphi \rightarrow [\delta _G] K_i \varphi$	(perfect recall)	(PR)
$([\delta _G] \varphi \wedge [\delta' _H] \psi) \rightarrow [\delta _G \cup \delta' _H] (\varphi \wedge \psi)$	(superadditivity)	(S)
(if $G \cap H = \emptyset$)		
From φ and $\varphi \rightarrow \psi$ infer ψ	(modus ponens)	(MP)
From φ infer $K_i \varphi$	(necessitation for knowledge)	(NK)
From φ infer $[\delta _G] \varphi$	(necessitation for action)	(NA)

We now turn our attention to some important CEDL validities, displayed in Table 1.

Theorem 1 *The principles in Table 1 are sound and complete with respect to the class of CEDL models.*³

The axiomatisation in Table 1 reveals one additional assumption that is implicit in CEDL models. This assumption corresponds to Axiom (S), which is called ‘super-additivity’. It stipulates that if group G obtains outcome φ by acting as determined by the partial joint action δ and H (disjoint from G) obtains outcome ψ by acting as determined by the partial joint action δ' , then the group of agents $G \cup H$ obtains outcome $\varphi \wedge \psi$ by acting as determined by the union of their partial joint actions. In particular, this implies that the bigger the group, the more it can achieve. This seems to be an intuitive property.

2.3 Group Knowledge

To formalise collective responsibility, we have to provide a formalisation of ‘group knowledge’. Unfortunately, there is no consensus in the literature of what group knowledge means (see, e.g., the discussion in [18]). Here, we choose to use the notion of ‘distributed knowledge’, found, e.g., in [17]. In the language, we replace operators K_i by the more general K_G , which semantics is formally defined as follows. Let $G \neq \emptyset$:

$$M, w \models K_G \varphi \text{ iff } M, w' \models \varphi \text{ for all } w' \in \bigcap_{i \in G} \mathcal{K}(i)(w)$$

³For the proof we refer to [12].

Distributed knowledge approximately describes the knowledge of someone who has complete knowledge of what each member of the community knows. For example, assume that agent i knows that p is true but does not know whether q is true or false, whereas agent j knows that q is true but does not know whether p is true or false. Then, the group $\{i, j\}$ has distributed knowledge that $p \wedge q$ is true.

Proposition 1 *The following schemata are valid in CEDL with distributed knowledge:*

$$\begin{array}{lll}
 (K_G \varphi \wedge K_G(\varphi \rightarrow \psi)) \rightarrow K_G \psi & (\text{deductive closure for knowledge}) & (\text{KK}') \\
 K_G \varphi \rightarrow \varphi & (\text{truth}) & (\text{T}') \\
 K_G \varphi \rightarrow K_G K_G \varphi & (\text{positive introspection}) & (4') \\
 \neg K_G \varphi \rightarrow K_G \neg K_G \varphi & (\text{negative introspection}) & (5') \\
 K_G[\delta|_N] \varphi \rightarrow [\delta|_N] K_G \varphi & (\text{perfect recall}) & (\text{PR}') \\
 (K_{G_1} \varphi_1 \wedge K_{G_2} \varphi_2) \rightarrow K_{G_1 \cup G_2} (\varphi_1 \wedge \varphi_2) & (\text{knowledge superadditivity}) & (\text{KS}) \\
 (\text{if } G \cap H = \emptyset) & &
 \end{array}$$

Proof KK' to $5'$ are easy, and left for the reader. PR' is true because (C2) is preserved for groups G by the definition of K_G . And KS is true because $\bigcap_{i \in G_1 \cup G_2} \mathcal{K}(i) \subseteq \bigcap_{i \in G_1} \mathcal{K}(i)$. $\square$

Note that with axioms KS and KK' , one can derive both: $K_{G_1} \varphi \rightarrow K_{G_1 \cup G_2} \varphi$ and $(K_{G_1} \varphi_1 \wedge K_{G_2} (\varphi_1 \rightarrow \varphi_2)) \rightarrow K_{G_1 \cup G_2} \varphi_2$. However, whether these principles constitute a complete axiomatic system for CEDL with distributed knowledge is left as an open question.

2.4 Comparison with CEGMs

The models of CEDL are very similar to *Concurrent Epistemic Game Models* (CEGMs), structures which are commonly used to model joint action of epistemic agents (see [1, 9, 16, 24, 25] for some examples). There are four principle differences, one of which is, from our perspective, significant.

- CEGMs do not necessarily satisfy (C2). This is not a deep difference, since we could either relax (C2) for CEDLs or require it for CEGMs.
- CEDL models do not necessarily satisfy *Determinism* (DET) (the outcome of full joint actions is deterministic, i.e. a singleton or empty), which CEGMs require.⁴ This is again not deep, for the same reasons as above.
- CEDL models do not necessarily satisfy *Action Awareness* (AA), the principle that if two states are indistinguishable for i , then the same actions are executable for i in each state. Again, we have the options of relaxing this condition for

⁴Non-deterministic CGMs (CEGMs without the epistemic features, which can be easily added) are discussed in [7].

- CEGMs or requiring it for CEDLs. See Example 3 for a CEDL model which does not satisfy AA.
- CEGMs, by their very specification, require a principle of *Independent Choice* (IC) [2], that is, that if an action a is executable by agent i in a given world w , then it is executable *no matter what the other agents do*. This is baked into the usual presentation of CEGMs (or, more generally, CGMs), since one specifies the individual actions i can do in w and then builds up joint actions from these. On the other hand, CEDL models start with joint actions and require no such independence principle.

Condition (IC) aside, the differences between CEGMs and CEDL models are not essential. The conditions (C2), (DET) and AA are requirements imposed in addition to the basic structure of these models, whereas (IC) occurs in as a fundamental step in the construction of CEGMs. We will argue in Section 5.1 that (IC) is unsatisfactory for our needs: the most obvious way to enforce (IC) on CEDL models entails an unreasonable consequence for accountability. On the other hand, weakening (DET) and AA or requiring C2 for CEGMs is not as essential to their usual definition.

It is not hard to show that, fixing the sets P of propositions, N of agents and A of actions, the category of CEGMs satisfying (C2) is isomorphic to the category of CEDL models satisfying (DET), AA and (IC). The proof is not difficult, but long and fairly tedious and so we omit it here.

3 Ability and Knowing How

Our aim in this section is to define operators to express agent ability. The first operator we define here expresses that “by executing $\delta|_G$, the group G of agents ensures an outcome satisfying φ in the next step”. That is, formulae of the form $E_{\delta|_G}\varphi$ should mean that G *can* execute $\delta|_G$ (i.e. $\delta|_G$ is executable) and that it necessarily leads to an outcome satisfying φ . Thus, we define $E_{\delta|_G}$ as follows:

$$E_{\delta|_G}\varphi := \neg[\delta|_G]\perp \wedge [\delta|_G]\varphi.$$

Sometimes, we will need to express that some group of agents ensure an outcome φ by executing a sequence of actions $\delta_1|_G; \dots; \delta_n|_G$. Hence, we define

$$E_{\delta_1|_G; \dots; \delta_n|_G}\varphi := \neg[\delta_1|_G; \dots; \delta_n|_G]\perp \wedge [\delta_1|_G; \dots; \delta_n|_G]\varphi.$$

The following properties relate the group and subgroup ability and are used in proofs of theorems which follow.

Lemma 1 *Let $G' \subseteq G$. The following are theorems of CEDL.*

1. $[\delta_1|_{G'}; \dots; \delta_n|_{G'}]\varphi \rightarrow [\delta_1|_G; \dots; \delta_n|_G]\varphi$
2. $\neg[\delta_1|_G; \dots; \delta_n|_G]\perp \rightarrow ([\delta_1|_{G'}; \dots; \delta_n|_{G'}]\varphi \rightarrow E_{\delta_1|_G; \dots; \delta_n|_G}\varphi)$
3. $\exists \delta|_{G'} \langle \delta|_{G'} \rangle \varphi \rightarrow \exists \delta|_G \langle \delta|_G \rangle \varphi$
4. $(E_{\delta_1|_G; \dots; \delta_n|_G}\varphi \wedge E_{\delta_1|_G; \dots; \delta_n|_G}(\varphi \rightarrow \psi)) \rightarrow E_{\delta_1|_G; \dots; \delta_n|_G}\psi$
5. $E_{\delta_1|_{G'}; \dots; \delta_n|_{G'}}\varphi \rightarrow \exists \delta'_1|_G; \dots; \delta'_n|_G \succ \delta_1|_{G'}; \dots; \delta_n|_{G'} \varphi \rightarrow E_{\delta'_1|_G; \dots; \delta'_n|_G}\varphi$

There is a technical detail regarding our definition of E which will require some attention. While

$$\models E_{\delta|_G} E_{\delta_1|_G; \dots; \delta_n|_G} \varphi \rightarrow E_{\delta|_G; \delta_1|_G; \dots; \delta_n|_G} \varphi,$$

the converse is not true in general. This is because $E_{\delta|_G} E_{\delta_1|_G; \dots; \delta_n|_G} \varphi$ entails that

$$\neg[\delta|_G]\perp \wedge [\delta|_G]\neg[\delta_1|_G; \dots; \delta_n|_G]\perp, \quad (1)$$

and this implies, but is not implied by,

$$\neg[\delta|_G][\delta_1|_G; \dots; \delta_n|_G]\perp. \quad (2)$$

Condition (1) requires that $\delta_1|_G; \dots; \delta_n|_G$ is executable in every world reachable by doing $\delta|_G$, whereas (2) requires only that there is some world w reachable by doing $\delta|_G$ such that $\delta_1|_G; \dots; \delta_n|_G$ is executable from w (see Fig. 3).

This difference will complicate some of the statements of our theorems. It would be possible to avoid these complications, of course, by defining $E_{\delta|_G; \delta_1|_G; \dots; \delta_n|_G}$ to be $E_{\delta|_G} E_{\delta_1|_G; \dots; \delta_n|_G}$. Nonetheless, there are good reasons to define E as we have. It is common to define so-called “test actions” $\varphi?$ in dynamic logics, where

$$\mathcal{T}_{\varphi?}(w) = \begin{cases} \{w\} & \text{if } w \models \varphi, \\ \emptyset & \text{else.} \end{cases}$$

In the presence of such actions, it is common to hit a “dead end” along some paths and not others, and in that context, our definition of E is more appropriate than the alternative. While we do not explicitly include test actions in our presentation, they are a natural addition to our model, and so we shouldn’t presume that every path can be executed all the way to the end of $\delta_1|_G; \dots; \delta_n|_G$.

Let us call a sequence $\delta_1|_G; \dots; \delta_n|_G$ *uniform at w* just in case $M, w \models U_{\delta_1|_G; \dots; \delta_n|_G}$, where

$$U_{\delta_1|_G; \dots; \delta_n|_G} = E_{\delta_1|_G} \dots E_{\delta_n|_G} \top.$$

In Fig. 3, $\delta_1|_G; \delta_2|_G$ is uniform at w'_0 , but not at w_0 .

The following properties are useful for some of the proofs hereafter. The second is used in the proof of Theorem 3 and the third property is proved using the derived rule of inference

$$\text{From } \varphi \rightarrow \psi, \text{ infer } E_{\delta_1|_G; \dots; \delta_n|_G} \varphi \rightarrow E_{\delta_1|_G; \dots; \delta_n|_G} \psi.$$

Fig. 3 Both w_0 and w'_0 satisfy $E_{\delta_1|_G; \delta_2|_G} \top$, but $E_{\delta_1|_G} E_{\delta_2|_G} \top$ is true at w'_0 and not at w_0

Lemma 2 *The following are theorems of CEDL.*

1. $U_{\delta_1|G;\dots;\delta_n|G} \rightarrow E_{\delta_1|G} U_{\delta_2|G;\dots;\delta_n|G}$
2. $KG U_{\delta_1|G;\dots;\delta_n|G} \rightarrow E_{\delta_1|G} KG U_{\delta_2|G;\dots;\delta_n|G}$
3. $(U_{\delta_1|G;\dots;\delta_n|G} \wedge E_{\delta_1|G;\dots;\delta_n|G} \varphi) \leftrightarrow E_{\delta_1|G} \dots E_{\delta_n|G} \varphi.$

With this definition of ability in hand, we turn our attention to know-how. There is a subtle distinction to be made here: one may know that there is a way to realise φ , but not know what the strategy is. In other words, she knows *that* φ is realisable, but she does not know *how* to realise φ . We can see this situation in our Light Bulb and Light Switch scenario (Example 2): let us recall that, at w_0 , Betty is able to ensure that the light is on after one step. In fact, this is true at w_1 too. Hence, at w_0 , Betty knows that she can ensure p in one step, but Betty does *not* know that the action *tog* ensures p . Thus, Betty knows that she is able to ensure that the light is on after one step, but she does not know what she must do to ensure it! In game theory [32] we say that an agent has a “non-uniform⁵ strategy” for a given goal whenever

for every state that the agent cannot distinguish from the current state, there is a strategy whose execution leads to the goal;

and we say that an agent has a “uniform strategy” for a given goal whenever

there is a strategy such that for every state that the agent cannot distinguish from the current one, its execution leads to the goal.

In terms of our logic, agent i has a non-uniform strategy for φ at w just in case

$$M, w \models K_i \exists_{\delta \in \Delta} E_{\delta|_i} \varphi,$$

whereas she has a uniform strategy for φ at w if and only if

$$M, w \models \exists_{\delta \in \Delta} K_i E_{\delta|_i} \varphi.$$

This leads naturally to a definition of know-how for groups. A group G knows how to realise φ just in case there is a non-empty subgroup G' of G such that (1) G' can realise φ by performing a sequence of actions $\delta_1|_{G'}; \dots; \delta_n|_{G'} \varphi$ and (2) each member of G' knows that $\delta_1|_{G'}; \dots; \delta_n|_{G'} \varphi$ will realise φ . In this case, no matter what other members of the group do, each of the members of G' knows how G' can ensure that φ . Consequently, we define

$$H_G \varphi := \exists_{\emptyset \neq G' \subseteq G} \exists_{\delta \in \Delta} C_{G'} E_{\delta|_{G'}} \varphi,$$

where

$$C_G \varphi := \forall_{i \in G} K_i \varphi$$

represents that φ is *universal knowledge* for the group G .

In the following proposition and hereafter, H_G^n is a sequence of n operators H_G . Explicitly,

$$\begin{aligned} H_G^0 \varphi &= \varphi, \\ H_G^{n+1} \varphi &= H_G H_G^n \varphi \end{aligned}$$

⁵This use of the term “uniform” is unrelated to our definition of U above.

As the proposition shows, $H_G^n \varphi$ expresses the claim that some part of group G knows how to ensure φ in n steps. In the case that $n = 0$, the claim is trivial, since there is nothing for the group to know how to do — either φ attains or not.

Proposition 2 *The following are valid in CEDL.*

1. $\neg H_G \perp$
2. $H_G^n \varphi \rightarrow \exists \emptyset \neq G' \subseteq G \exists \delta_1|_G; \dots; \delta_n|_G C_{G'} E_{\delta_1|_{G'}} \dots E_{\delta_n|_{G'}} \varphi$
3. *From $\varphi \rightarrow \psi$, infer $H_G \varphi \rightarrow H_G \psi$.*

4 Responsibility-as-obligation

In this section, we define an operator expressing a group's obligations, as in deontic logic (see [29]). We do so by adapting the simple, yet effective, idea of Anderson and Moore [4] to our framework, rewritten in [11]. The idea is that “ φ is obligatory if and only if failure to do what is required to make φ true would lead to sanction.” For example, if it is obligatory for some agent i that the light is on, then every possible world where it is false satisfies the proposition that the agent i is sanctioned. The way the static and dynamic obligations are introduced in PDL in d'Altan et al. is based on ideas of Meyer [28] which is important inspiration for our treatment of the subject.

From now on, we assume that the set of propositional variables of our language is $P \cup \text{Vio}$, where $\text{Vio} = \{\text{vio}_G \mid G \in 2^N \setminus \emptyset\}$. That is, for each group of agents, we introduce a new atomic formula vio_G , representing the condition that “the group of agents G is in violation.” In addition, we require that our models now be quadruples of the form $\langle W, \mathcal{K}, \mathcal{T}, \mathcal{V} \rangle$, where W , $\mathcal{K}$ and $\mathcal{T}$ are defined as before but the domain of $\mathcal{V}$ is extended to the new set of propositional variables, i.e., $\mathcal{V} : (P \cup \text{Vio}) \rightarrow 2^W$.

In addition, we introduce an operator O_G , which expresses that a particular state of affairs φ is obligatory, equivalently, that $\neg\varphi$ entails a violation. Hence, we define

$$O_G \varphi := \neg\varphi \rightarrow \text{vio}_G.$$

Obligations satisfy some interesting properties.

Proposition 3 *The following theorems and rule of inference are valid in CEDL.*

1. $O_G \top$
2. $O_G(\varphi \wedge \psi) \leftrightarrow (O_G \varphi \wedge O_G \psi)$
3. *From φ , infer $O_G \varphi$.*

We note that agents do not necessarily know their obligations, i.e., $O_G \varphi$ does not imply $K_G O_G \varphi$. This is consistent with traditional views of morality, dating back to Aristotle's *Nicomachean Ethics*.

Also, our approach permits models where violations are unavoidable. For instance, we have that $\text{vio}_G \rightarrow O_G \perp$ is valid, which also means that we do not impose Axiom D (i.e., $\neg O_G \perp$), which is present in some deontic systems. In our view it just means that dilemmas are possible (cf. [40]). As in Sartre's classic example, one can have the

obligation to stay at home to look after an elderly mother and, at the same time, have an obligation to join the resistance movement to fight the Nazis.

With this definition of obligation we formalize one kind of forward-looking responsibility, namely, responsibility-as-obligation: the “obligation to see to it that”. The aim here is to augment our logic with operators R , where formulae of the form $R_G^{n+1}\varphi$ are to be read as “it is obligatory for G to see to it that φ is true after n steps.” Goodin’s ideas are used to define this meaning of responsibility. The argument advanced is the following quote from Goodin ([19], p. 51):

The standard form of responsibility is that A *see to it that* φ . It is not enough that φ occurs. A must have “seen to it” that φ occurs. “Seeing to it that φ ” requires, minimally: that A satisfy himself that there is some process (mechanism or activity) at work whereby φ will brought about; that A check, from time to time, to make sure that that process is still at work, and is performing as expected; and that A take steps as necessary to alter or replace processes that no longer seem likely to bring about φ .

This is a complex definition. It is unlikely that we will be able to capture all its details in our framework. Nonetheless, we may approximate it by the following inductive definition:

$$\begin{aligned} R_G^0\varphi &:= O_G H_G^0\varphi, \\ R_G^{n+1}\varphi &:= O_G H_G^{n+1}\varphi \wedge \forall_{\delta \in \Delta} [\delta] R_G^n\varphi. \end{aligned}$$

The formula $R_G^0\varphi$ is equivalent to $O_G\varphi$. That is, $R_G^0\varphi$ is true at some world w if and only if $\neg\varphi$ implies a violation for G . But formulae $R_G^n\varphi$, for $n > 1$, are more than mere obligations that one does what is necessary to ensure φ . They also include the “supervisory nature” of this kind of responsibility, because they also require that, after n steps, $\neg H_G\varphi$ implies a violation for G . Intuitively, the group is responsible for discovering a plan, that is, for knowing how to accomplish their task of ensuring φ . As a result, members of the group may be responsible for sharing the requisite knowledge to ensure that the group actually *does* know how to realise φ .

Lemma 3 *The following theorem is valid in CEDL.*

$$R_G^{n+1}\varphi \rightarrow R_G^1 H_G^n\varphi.$$

Proof

$$\begin{aligned} R_G^{n+1}\varphi &\Leftrightarrow O_G H_G^{n+1}\varphi \wedge \forall_{\delta \in \Delta} [\delta] R_G^n\varphi \\ &\Leftrightarrow O_G H_G H_G^n\varphi \wedge \forall_{\delta \in \Delta} [\delta] (O_G H_G^n\varphi \wedge \forall_{\delta' \in \Delta} [\delta'] R_G^n\varphi) \\ &\Rightarrow O_G H_G H_G^n\varphi \wedge \forall_{\delta \in \Delta} [\delta] O_G H_G^n\varphi \\ &\Leftrightarrow O_G H_G H_G^n\varphi \wedge \forall_{\delta \in \Delta} [\delta] R_G^0 H_G^n\varphi \\ &\Leftrightarrow R_G^1 H_G^n\varphi. \quad \square \end{aligned}$$

In the next two sections, we will discuss two meanings of backward-looking responsibility, namely responsibility as accountability and responsibility as

blameworthiness. These responsibilities are backward-looking in the sense that they usually apply to something that has occurred.

5 Accountability

Let us start by recalling the conditions for “accountability” given in [39]:

... an agent i is accountable for φ if i has the capacity to act responsibly (has moral agency), is somehow causally connected to the outcome (by an action or omission) and there is a reasonable suspicion that agent i did somehow something wrong.

The first condition for holding an agent accountable, “moral agency”, is a tacit assumption in our framework. In fact, our logical framework also assumes that all agents are rational and that they are perfect reasoners. This, for instance, means that agents can foresee all the logical consequences of the facts they know. Again, these assumptions may be seen as very restrictive ones, but, in a logical framework like ours, the relaxation of such assumptions would make it much more difficult (if not impossible) to derive some interesting properties.

The other two conditions for holding an agent accountable are causality and wrong-doing. In the context of a responsibility to see to it that $\neg\varphi$, these two conditions must be related to each other, in the sense that being causally connected to a prohibited outcome is itself “something wrong” done by the agent. There are two distinct ways in which the agent’s action can be causally connected to the outcome.

The more direct way is that the agent did something that ensured φ when $\neg\varphi$ was obligatory. More explicitly, the actions that the agent took necessarily resulted in φ , no matter what others did. In this case, the agent is accountable for the fact that φ . We can characterize this condition by introducing a relation $\mathcal{A}^E$ which expresses that the agent’s (or group’s) actions have necessitated φ . Let G be a group such that $R_G^n \neg\varphi$ and let $G' \subseteq G$. Then G' is *directly accountable* for φ by doing $\delta_1|_{G'}; \dots; \delta_n|_{G'}$ iff $\mathcal{A}_{\delta_1|_{G'}; \dots; \delta_n|_{G'}}^E \varphi$ holds where

$$\begin{aligned}\mathcal{A}_{\delta|_{G'}}^E \varphi &:= E_{\delta|_{G'}} \varphi, \\ \mathcal{A}_{\delta_1|_{G'}; \dots; \delta_{n+1}|_{G'}}^E \varphi &:= E_{\delta_1|_{G'}} \mathcal{A}_{\delta_2|_{G'}; \dots; \delta_{n+1}|_{G'}}^E \varphi.\end{aligned}$$

Thus, G' is directly accountable for φ by doing $\delta_1|_{G'}; \dots; \delta_n|_{G'}$ iff $E_{\delta_1|_{G'}} \dots E_{\delta_n|_{G'}} \varphi$, that is, iff doing $\delta_1|_{G'}; \dots; \delta_n|_{G'}$ will realise φ *no matter what the rest of the group G does*. In this case, G' has really brought about φ all by themselves, so to speak.⁶

An agent (or subgroup) can also be indirectly accountable for φ by failing to ensure that the group as a whole knows how to avoid φ . This can happen in two distinct

⁶To be sure, it may be the case that φ was inevitable, in which case *every* subgroup of G is unfortunately accountable for φ , no matter what (executable) $\delta_1|_G; \dots; \delta_n|_G$ was performed! In a no-win situation, everyone is accountable for the failure. This is not as unfair as it sounds, since while they are accountable, they are not blameworthy (see below).

ways. First, it may be that i could have ensured that the group G knows how to avoid φ if only i had shared some of his knowledge—for instance, if Alice fails to tell Betty that the light is off, then she is indirectly accountable in this sense. In such a case, if the actions the group actually pursues fail to realise φ , then i is accountable for her failure to share information. On the other hand, it could be that the group currently knows how to realise $\neg\varphi$, but the agent does something unexpected, no one in the group knows how to recover (though recovery is possible) and instead realises φ . Again, in this case, the agent is accountable for the group's failure to realise φ . Thus, in the same conditions as above, G' is *indirectly accountable* for φ iff $\mathcal{A}_{\delta_1|_{G'}; \dots; \delta_n|_{G'}}^H \varphi \wedge E_{\delta_1|_G; \dots; \delta_n|_G} \varphi$ where

$$\begin{aligned} \mathcal{A}_{\delta|_{G'}}^H \varphi &:= \perp, \\ \mathcal{A}_{\delta_1|_{G'}; \dots; \delta_{n+1}|_{G'}}^H \varphi &:= E_{\delta_1|_{G'}} \mathcal{A}_{\delta_2|_{G'}; \dots; \delta_{n+1}|_{G'}}^H \varphi \vee (E_{\delta_1|_{G'}} \neg H_G^n \neg \varphi \wedge \exists \delta|_{G'} (\delta|_{G'}) H_G^n \neg \varphi). \end{aligned}$$

Note that the base case is trivial. This is because if one has a one-step responsibility that $\neg\varphi$, there is no responsibility regarding group knowledge in the next step. More explicitly, in the case that $n = 0$, Lemma 3 reduces to a tautology,

$$R_G^1 \varphi \rightarrow R_G^1 \varphi,$$

and so for one-step responsibilities, any accountability must be direct. The second disjunct in the inductive step requires that, if some subgroup G' is in a situation in which G' could have ensured that the group G had the know-how to ensure $\neg\varphi$, but G' 's action instead ensured that G hadn't the requisite know-how, then G' is accountable for this failure. Recall, of course, that when we say a group G has the know-how to ensure $\neg\varphi$, what we mean is that some subgroup of G has the know-how to do so. This is sufficient to ensure $\neg\varphi$.

We can now define accountability by merging these two conditions. Then a subgroup $G' \subseteq G$ is accountable for φ by doing $\delta_1|_{G'}; \dots; \delta_n|_{G'}$ iff $\mathcal{A}_{\delta_1|_{G'}; \dots; \delta_n|_{G'}} \varphi$, where

$$\begin{aligned} \mathcal{A}_{\delta_1|_{G'}; \dots; \delta_n|_{G'}} \varphi &:= R_G^n \neg \varphi \wedge \\ &(\mathcal{A}_{\delta_1|_{G'}; \dots; \delta_n|_{G'}}^E \varphi \vee \\ &(\mathcal{A}_{\delta_1|_{G'}; \dots; \delta_n|_{G'}}^H \varphi \wedge E_{\delta_1|_G; \dots; \delta_n|_G} \varphi)). \end{aligned}$$

Note that the inclusion of indirect accountability (i.e. $\mathcal{A}^H$) is an important feature of accountability involving group coordination. A subgroup is accountable if it could have ensured that some portion of the larger group could know how to bring about $\neg\varphi$ but failed to convey the relevant information. Accountability is not just about what one does in pursuit of $\neg\varphi$, but also about the information one could make available (chiefly, though not exclusively, by communication, as in our lightbulb example). This aspect of accountability, which was also present as the “supervisory nature” condition in our account of forward-looking responsibility, is missing from other accounts of group accountability and blameworthiness, including the models found in [10] and [27], discussed in Section 6. As far as we know, this is an original feature of our model.

The following theorem confirms that our definition of accountability satisfies some of our most basic intuitions.

Theorem 2 *Let $G' \subseteq G'' \subseteq G$. The following are theorems of CEDL.*

1. $R_G^n \neg\varphi \rightarrow (\mathcal{A}_{\delta_1|_{G'}; \dots; \delta_n|_{G'}}^E \varphi \rightarrow E_{\delta_1|_{G'}; \dots; \delta_n|_{G'}} \varphi)$
2. $U_{\delta_1|_{G''}; \dots; \delta_n|_{G''}} \rightarrow (\mathcal{A}_{\delta_1|_{G'}; \dots; \delta_n|_{G'}} \varphi \rightarrow \mathcal{A}_{\delta_1|_{G''}; \dots; \delta_n|_{G''}} \varphi)$

Item (1) says that a (sub)group is directly accountable for φ only if their actions necessitated φ , while Item (2) says that accountability propagates upward, assuming that the sequence $\delta_1; \dots; \delta_n$ is uniform and executable by G'' .

Example 2 (Light Bulb and Light Switch with Violations (revisited)) To illustrate operator $\mathcal{A}$, we reuse the scenario of Example 1 and the model in Fig. 2. The only difference in this variation is that worlds w_0 , w_1 and w_2 are violation states for the group. This scenario is depicted in Fig. 4.

The careful reader may wonder why w_1 is a violation state, given that $M, w_1 \models p$. In fact, this is a little “hack” to ensure that $M, w_0 \models R_G^2 p$, as desired. By Lemma 3 and the definition of $R_G^1 p$, this requires that every world reachable in one step from w_0 satisfies $O_G H_G^1 p$. Since w_1 does *not* satisfy $H_G^1 p$, we must set it to a violation state. One can confirm that, in this model, w_0 does indeed satisfy $R_G^2 p$ (and, in fact, so do each of the other worlds in M).

One can confirm the following.

- $M, w_0 \models \mathcal{A}_{\delta_0|_a; \delta_0|_a} \neg p$: In particular, it is easy to see that Alice is indirectly accountable for $\neg p$.
- $M, w_0 \models \neg \mathcal{A}_{\delta_2|_a; \delta_0|_a} \neg p$: Theorem 2(1) entails that Alice is not directly accountable in this case, and a careful examination shows that neither she is indirectly accountable.
- $M, w_0 \models \mathcal{A}_{\delta_0|_b; \delta_0|_b} \neg p$: Betty is directly accountable, since $M, w_0 \models E_{\delta_0|_b} E_{\delta_0|_b} \neg p$.

Fig. 4 The light bulb and switch scenario, with dotted circles indicating violation states

- $M, w_0 \models \mathcal{A}_{\delta_0|_{\{a,b\}}; \delta_0|_{\{a,b\}}} \neg p$: Follows immediately from Theorem 2(2).
- $M, w_0 \models \mathcal{A}_{\delta_1|_{\{a,b\}}; \delta_3|_{\{a,b\}}} \neg p$: Direct accountability follows from $M, w_0 \models E_{\delta_1|_{\{a,b\}}; \delta_3|_{\{a,b\}}} \neg p$.

Finally, let us consider the situation in which the group executes $\delta_0|_G; \delta_1|_G$ in w_0 , resulting in the lucky happenstance that the light bulb ends up on, even though Betty did not know that this would be the case. Clearly, neither Alice nor Betty can be accountable for $\neg p$, because $\neg p$ was not realised. Nonetheless, Alice does *not* escape all accountability. Applying Lemma 3, we conclude that

$$M, w_0 \models R_G^1 H_G p,$$

and Alice is thereby accountable for $\neg H_G p$ after doing $\delta_1|_a$. She had a (one-step) responsibility to tell Betty the state of the bulb, and she neglected to do so.

Let us consider an alteration of this example. Add a third person, Cecile, also standing outside and able to toggle a second switch. The light will switch from on to off or vice versa iff exactly one of Betty and Cecile toggles her switch; if both toggle, then nothing happens. Cecile's knowledge relation is the same as Betty's, and both Cecile and Betty can see what action the other has taken. The amended graph is shown in Fig. 5.

As before, the group is responsible to reach a p state in two steps. If the group executes $\delta_0; \delta_0$, then they will fail to realise their obligation, and the group is accountable for their failure, i.e.

$$M, w_0 \models \mathcal{A}_{\delta_0|_G; \delta_0|_G} \neg p.$$

Moreover, Alice will also be (indirectly) accountable in this case, since she failed to notify Betty and Cecile of the state of the light.

But notice that neither Betty nor Cecile are accountable. Betty's actions did not necessitate $\neg p$ —had Cecile toggled her switch once, they would have realised p . And Cecile is not accountable either, for symmetric reasons.

However, the subgroup $\{b, c\}$ is accountable. The action sequence $\delta_0|_{\{b,c\}}; \delta_0|_{\{b,c\}}$ does indeed necessitate $\neg p$. Thus, we have an example of a subgroup which is accountable, though no individual in the subgroup is accountable.

Fig. 5 The light bulb and switch scenario, with a third participant. Both Betty and Cecile can toggle light switches, which act as an XOR gate

Thus, our model satisfies the following.

- $M, w_0 \models \mathcal{A}_{\delta_0|a;\delta_0|a} \neg p$: If Alice does nothing in either step, then she is (indirectly) accountable, since she failed to alert Betty regarding the state of the light.
- $M, w_0 \models \neg \mathcal{A}_{\delta_4|a;\delta_0|a} \neg p$: Alice alerted Betty in the first step, but Betty and Cecile did nothing. Alice is not accountable for the failure.
- $M, w_0 \models \neg \mathcal{A}_{\delta_4|b;\delta_0|b} \neg p$: Betty is not accountable for the failure to turn the light on, since her actions did not necessitate that the light was off after two steps. Cecile could have turned on the light as Betty did nothing, after all, ending in a desired state.
- $M, w_0 \models \mathcal{A}_{\delta_0|[b,c];\delta_0|[b,c]} \neg p$: The subgroup consisting of Betty and Cecile are accountable for $\neg p$, even though neither of the two are individually accountable.

5.1 Accountability and Independent choice

Let us return to the topic of independent choice discussed in Section 2.4 (see also [2]), that is, the property that, if an agent i can perform an action in a in world w , then the agent can perform that action no matter what other agents do. Formally, this amounts to

$$\neg[\delta|_G]\perp \rightarrow \neg[\delta]\perp.$$

Intuitively, this is an implausible condition, especially where cooperation is required for certain joint actions. It is possible that Alice can execute α iff Betty executes β . Imagine Alice needs a boost to get over a fence and only Betty can provide that boost. Then Alice can perform *scale* iff Betty performs *boost*, thus violating (IC).

One might think that (IC) is a matter of perspective. After all, one might interpret the action *scale* as “Alice attempts to scale the fence.” Under this interpretation, *scale* would be executable in every world but makes little or no difference in worlds where Betty does not cooperate. However, this interpretation leads to unfortunate consequences for accountability (and hence blameworthiness).

Consider the two situations illustrated in Fig. 6. The lefthand figure represents a model in which Alice can perform *scale* iff Betty chooses to *boost*. This model obviously does not satisfy (IC). The righthand model is one in which we interpret *scale* in terms of *attempting* to cross the fence. If Betty doesn’t cooperate, then Alice’s attempt to cross the fence results in no change to the world. Suppose, further, that the group has an obligation to ensure Alice does not cross the fence (in one step).

However, notice that in w_0 , Alice is directly accountable for crossing the fence by executing *scale*. This is because every execution of *scale* from w_0 ends in a violation state (w_1). In w'_0 , however, while the group $\{a, b\}$ is accountable for crossing the

Fig. 6 The left-hand diagram shows a natural CEDL model for scaling a fence, while the right-hand shows the most plausible CEDL model satisfying Independent Choice

fence if they execute the joint action $\{a : \text{scale}, b : \text{boost}\}$, neither Alice nor Betty is accountable for her individual actions. This does not seem plausible. The action *scale* aims at reaching w'_1 and will do so, as long as Betty cooperates. There is, indeed, no other reason to execute *scale* (attempting to scale the fence is a poor substitute for doing nothing!). An appropriate model would judge that Alice is accountable for choosing *scale*, Betty accountable for choosing *boost* and the group accountable for the joint action.

For this reason, we regard (IC) as inappropriate for the notion of accountability as we've sketched it here and that CEDL models provide a more appropriate semantics for our theory than CEGMs.

6 Blameworthiness

The definition given in [39] for this meaning of responsibility is the following: agent i is blameworthy for consequence φ whenever i is accountable for φ and is not capable to give an acceptable account for it. Two accounts (excuses) are considered acceptable. The first one is ignorance: the agent i is excused (and thus is not blameworthy) for φ if i can show that she could not know that her behavior would cause φ . The second is compulsion. A possible interpretation of compulsion is that the agent i is excused for φ if i can show that no behavior that does not cause a violation is possible.

We extend van de Poel's definition to include groups and their subgroups and not just individuals. Let $G' \subseteq G$ and suppose further that G has executed sequence of actions $\delta_1|_G; \dots; \delta_n|_G$. Define

$$\begin{aligned} \mathcal{B}_{\delta_1|_{G'}}\varphi &= K_{G'}\mathcal{A}_{\delta_1|_{G'}}\varphi \wedge \exists \delta'_1|_{G'} \langle \delta'_1|_{G'} \rangle \neg \text{vio}_G \\ \mathcal{B}_{\delta_1|_{G'}; \dots; \delta_{n+1}|_{G'}}\varphi &= (R_G^{n+1} \neg \varphi \wedge E_{\delta_1|_{G'}} \mathcal{B}_{\delta_2|_{G'}; \dots; \delta_{n+1}|_{G'}}\varphi) \\ &\quad \vee (K_{G'}\mathcal{A}_{\delta_1|_{G'}; \dots; \delta_{n+1}|_{G'}}\varphi \wedge \exists \delta'_1|_{G'}; \dots; \delta'_{n+1}|_{G'} \\ &\quad \langle \delta'_1|_{G'}; \dots; \delta'_{n+1}|_{G'} \rangle \neg \text{vio}_G) \end{aligned}$$

Note that the above definition uses distributed knowledge rather than common knowledge. If the group as a whole recognizes that a particular plan would result in accountability, then the group itself is blameworthy.

In both the base case and the inductive case, one can explicitly see that, if one knows she is accountable and is under no compulsion to end up in a violation state, then one is blameworthy. The inductive step also includes a clause ensuring that one is blameworthy if they are blameworthy at any stage in executing the action sequence. We discuss this feature in Example 4 below.

Clearly, our definition allows one to be accountable but not blameworthy if he lacked knowledge that his actions would make him accountable. Similarly, if one's actions necessitated $\neg\varphi$, but nothing that he could do would avoid a violation state, then he is accountable, but not blameworthy.

Our formalisation of blameworthiness is similar in spirit to *agentive responsibility* discussed in [27]. Both accounts lean heavily on Aristotle's remarks on voluntary actions [5] and blameworthiness (recast as a kind of responsibility in Lorini, et al.). For instance, our ignorance exception corresponds to the authors' *epistemic*

condition, while the compulsion exception corresponds to the *control condition*, at least in the passive sense. Lorini, et al., however, use atemporal group STIT in order to model agentive responsibility,⁷ which essentially amounts to focusing on blame for single actions rather than sequences of actions. Since organisations often must implement complicated plans in order to satisfy their obligations, we regard our framework as more appropriate for modeling blame within organisations.⁸

A very different approach to responsibility and blame can be found in [10], in which Chockler and Halpern develop a model based on causality. As a result, for instance, their definition of responsibility is broad enough that rain can be responsible (but not blameworthy) for the cancellation of a picnic, whereas our definitions of responsibility, accountability and blame apply only to agents and groups. There is also a difference insofar as the authors' definition of responsibility (backwards-looking and hence analogous to accountability in our terminology) includes the compulsion clause, whereas we reserve that clause for blameworthiness. One significant difference in Chockler and Halpern's account is that both responsibility and blame come in degrees, whereas on our account, one is either accountable (blameworthy, resp.) or not. Degrees of responsibility is a sensible feature, but one which has no obvious analog in our formal setting.

As was the case with accountability, we expect blame to propagate upwards. Because (distributed) knowledge is an essential part of blame, however, it will have to figure into the conditions that allow blame to apply to the supergroup of a blameworthy subgroup. The following example provides a situation where a subgroup is blameworthy, while the larger group is not, verifying the relevance of knowledge for upward propagation.

Example 3 Ethel and Fred are international spies currently being held by an international baddie. They are in a cell with two different doors, each with different lock mechanisms, and Fred has managed to lift a key from one of the guards. Unfortunately, neither Ethel nor Fred knows whether the key unlocks the left door or the right.

Their one chance at escape requires that Ethel distract the guards while Fred unlocks the proper door. As all international spies know, captured spies have a duty to attempt to escape. The situation is illustrated in Fig. 7. Ethel can either do nothing or distract the guard, while Fred can either unlock the left door or the right, depending on which key he has.⁹

⁷The authors also introduce definitions of guilt and reproach, quite different than our notion of blameworthiness. Rather, guilt and reproach are cast as an emotional reactions and hence dependent on the after-the-fact knowledge of the agent experiencing the emotion, whereas our sense of blameworthiness is intended as an objective normative status. Roughly, Lorini, et al., consider these emotions from an agent's perspective whereas we take a God's-eye view for blame.

⁸Single action responsibility also excludes consideration of indirect accountability discussed in the previous section, thus omitting the responsibility to communicate with fellow group members and the accountability if one fails to do so.

⁹As is common with such models, granularity of the actions yields different models. We could have used actions try_l and try_r rather than $unlock_l$ and $unlock_r$, but our aim here is to discuss actions which cannot be executed.

Fig. 7 A CEDL model for the escape scenario. Both Ethel and Fred have the same knowledge and p represents escape from the international baddie

Suppose that the actual world is w_0 , where Fred has the key to the righthand door. In this world, the group $\{e, f\}$ is responsible for p (escaping the cell) in one step, i.e.

$$M, w_0 \models R_G^1 p.$$

We can confirm the following.

- $M, w_0 \models \mathcal{A}_{\delta_0|e} \neg p$: If Ethel does nothing, it is certain that the escape will fail and she is thereby accountable.
- $M, w_0 \models \neg \mathcal{A}_{\delta_0|f} \neg p$: If Fred tries to unlock the right door, then he is not accountable for failure to escape, since he would have succeeded had Ethel distracted the guards.
- $M, w_0 \models \mathcal{A}_{\delta_0|\{e,f\}} \neg p$: The group as a whole would be accountable for failure to escape if Ethel does nothing, because accountability propagates upward.
- $M, w_0 \models \mathcal{B}_{\delta_0|e} \neg p$: Ethel would be blameworthy for doing nothing, since she knows the plan would not work unless she distracts the guards.
- $M, w_0 \models \neg \mathcal{B}_{\delta_0|\{e,f\}} \neg p$: The group as a whole is not blameworthy if Ethel does nothing, since the group does not know that $\delta_0|\{e,f\}$ is executable, even though $\delta_0|e$ is known to be executable.

The point here is that, in w_0 , Ethel is blameworthy if she does nothing, because she knows that she will be accountable in that case. The group is *not* blameworthy for doing δ_0 , because the group does not know that this is executable and hence don't know that they will be accountable.

The following theorem gives explicit conditions that allow one to conclude that a supgroup is blameworthy on the basis that a subgroup is blameworthy.

Theorem 3 Let $G' \subseteq G'' \subseteq G$. The following are theorems of CEDL.

1. $\mathcal{B}_{\delta_1|G; \dots; \delta_n|G} \varphi \rightarrow \mathcal{A}_{\delta_1|G; \dots; \delta_n|G} \varphi$
2. $K_{G''} U_{\delta_1|G''; \dots; \delta_n|G''} \rightarrow (\mathcal{B}_{\delta_1|G'; \dots; \delta_n|G'} \varphi \rightarrow \mathcal{B}_{\delta_1|G''; \dots; \delta_n|G''} \varphi).$

Example 4 (Light bulb and switch re-revisited) Consider again the example of Alice and Betty, shown in Fig. 4.

First, let us take this opportunity to explain the disjunct

$$R_G^{n+1}\varphi \wedge E_{\delta_1|G}\mathcal{B}_{\delta_2|G';\dots;\delta_{n+1}|G'}\varphi \quad (3)$$

in our definition of $\mathcal{B}$. This clause ensures that, if there is an k such that

$$M, w \models E_{\delta_1|G} \dots E_{\delta_k|G}\mathcal{B}_{\delta_{k+1}|G';\dots;\delta_{n+1}|G'}\varphi,$$

that is, if $\delta_1|G; \dots; \delta_k|G$ necessarily leads to a state in which subgroup G' is blameworthy for the tail of the sequence $\delta_{k+1}|G'; \dots; \delta_{n+1}|G'$, then G' is blameworthy for the whole sequence $\delta_1|G'; \dots; \delta_{n+1}|G'$.

Our reasoning for this condition can be explained in terms of Alice and Betty. Consider the sequence $\delta_2; \delta_0$, where Alice alerts Betty regarding the state of the light, but Betty does not toggle the switch. In this case, Betty is accountable for $\neg p$, as we saw in Example 2. In addition, Betty will be blameworthy for executing δ_0 in the second step, that is,

$$M, w_2 \models \mathcal{B}_{\delta_0|b}\neg p.$$

But we also regard Betty as blameworthy for executing the whole sequence $\delta_2; \delta_0$. The fact that she is ignorant in w_0 is no excuse, since after one step, she knows what she ought to do. Ignorance at the beginning of the plan doesn't excuse knowingly failing to do what one ought later.

This is precisely the motivation for including the clause (3). Without it, we would conclude that the group $\{a, b\}$ is blameworthy for executing $\delta_2; \delta_0$, but neither a nor b are individually blameworthy. On the contrary, Alice did what she ought to have done and Betty did not, even though Betty knew better. Therefore, Betty is blameworthy here.

In addition, our model satisfies the following.

- $M, w_0 \models \mathcal{B}_{\delta_0|a;\delta_0|a}\neg p$: We must show that $M, w_0 \models K_a\mathcal{A}_{\delta_0|a;\delta_0|a}\neg p \wedge \exists \delta'_1|a;\dots;\delta'_2|a p$. The latter conjunct is, in fact, witnessed by $\delta_0|a; \delta_0|a$ itself, since it would be possible to end in w_3 if Alice did nothing at all (since Betty might toggle the switch on her on). Moreover, because $\mathcal{K}_a(w_0) = \{w_0\}$, Alice knows she will be accountable for executing $\delta_0|a; \delta_0|a$. Hence Alice is blameworthy.
- $M, w_0 \models \neg\mathcal{B}_{\delta_2|a;\delta_0|a}\neg p$. Obvious, since Alice is not accountable for $\delta_2|a; \delta_0|a$.
- $M, w_0 \models \neg\mathcal{B}_{\delta_0|b;\delta_0|b}\neg p$. While it is clear that there is a path to $\neg\text{vio}_G$, Betty is not blameworthy because she does not know that $\delta_0|b; \delta_0|b$ results in her accountability.
- $M, w_0 \models \mathcal{B}_{\delta_0|\{a,b\};\delta_0|\{a,b\}}\neg p$. Follows from Theorem 3(2) and Alice's blameworthiness.
- $M, w_0 \models \mathcal{B}_{\delta_1|\{a,b\};\delta_3|\{a,b\}}\neg p$. Follows from Theorem 3(2) and Alice's blameworthiness.

In the three-person variation (shown in Fig. 5), we can confirm the following. Since the reasoning is familiar, we leave it to the reader.

- $M, w_0 \models \mathcal{B}_{\delta_0|a;\delta_0|a}\neg p$
- $M, w_0 \models \neg\mathcal{B}_{\delta_4|a;\delta_0|a}\neg p$

- $M, w_0 \models \neg \mathcal{B}_{\delta_0|b; \delta_0|b} \neg p$
- $M, w_0 \models \neg \mathcal{B}_{\delta_0|[b,c]; \delta_0|[b,c]} \neg p$

6.1 Problem of many hands

The ‘problem of many hands’ (PMH) is a term coined by Thompson in [37], and studied in the context of organisations (e.g., by Bovens [8]). It is a problem that sometimes arises when we address collective responsibility: a group of individuals can reasonably be blamed for an undesirable outcome, while no member of the group can reasonably be blamed for the outcome. The core of this problem appears to be the potential gap between individual and collective responsibility. Following Van de Poel [39], we will use the following general characterisation of the PMH: The PMH occurs if a group G is blameworthy for φ , whereas none of the individuals of the group G is blameworthy for φ .

We can formalize this as follows:

$$PMH_{\delta_1|G; \dots; \delta_n|G} \varphi := \mathcal{B}_{\delta_1|G; \dots; \delta_n|G} \varphi \wedge \forall_i \neg \mathcal{B}_{\delta_1|i; \dots; \delta_n|i} \varphi$$

Formula $PMH_{\delta_1|G; \dots; \delta_n|G} \varphi$ means that by performing $\delta_1|G; \dots; \delta_n|G$, the problem of many hands arises in group G with respect to the outcome φ . Analysing the definitions of operator $\mathcal{B}$ a bit further we realise that the PMH may arise from two different sources:

1. the individuals do not have the ability to avoid the violation, or
2. the individuals do not have the necessary knowledge that they are accountable for the violation state.

An example of the PMH is the example of the three person variation of the light bulb and switch scenario (see Fig. 5). One can show that if Alice, Betty and Cecile execute the sequence $\delta_4|_{\{a,b,c\}}; \delta_3|_{\{a,b,c\}}$, then the group $\{a, b, c\}$ is blameworthy for the outcome $\neg p$, though none of the individuals is blameworthy.

Now it might be objected that if the three individuals are collectively blameworthy for the light being off, each of them is also individually *as member of the group* blameworthy. This argument, however, presupposes a stronger version of the reducibility thesis. This last thesis states that blameworthiness of a group can always be analysed in terms of individual blameworthiness (see [30]). The stronger version is formulated as follows (cf. [39]): if a group is blameworthy for φ , all individuals of the group are individually blameworthy, which can be formalised as follows:

$$\mathcal{B}_{\delta_1|G; \dots; \delta_n|G} \varphi \rightarrow \forall_{i \in G} \mathcal{B}_{\delta_1|i; \dots; \delta_n|i} \varphi$$

If we accept this strong version of reducibility, it follows that the PMH cannot occur. Another extreme position to avoid the PMH is that there is an individual who bears the ultimate responsibility for the actions of all the other individuals of the collective, such as the ministerial responsibility. So, if there is waste, corruption, or any other misbehaviour found to have occurred within a ministry, the minister is blameworthy

even if the minister had no knowledge of the actions. This ultimate responsibility can be formalized as follows:

$$\mathcal{B}_{\delta_1|_G; \dots; \delta_n|_G} \varphi \rightarrow \mathcal{B}_{\delta_1|_i; \dots; \delta_n|_i} \varphi, \text{ where } i \in G \text{ is the 'leader' of the group } G.$$

Work on multi-agent systems (MAS) tries to avoid the PMH by managing explicitly the interdependencies between the organisational activities [14], such as delegation activity (the flow of obligations with an organisation) and the information activity (the flow of knowledge with the organisation).

6.2 Short-term and long-term blame

In our definition of blameworthiness, we have interpreted Aristotle's notion of compulsion in terms of being unable to avoid a state of violation. In some ways, this is misleading, since Aristotle defines compulsion as follows [5]:

What is forced has an external principle, the sort of principle in which the agent or victim contributes nothing—if, for instance, a wind or human beings who control him were to carry him off.

This notion of compulsion *is* included in our definition of blame—if a wind carries one away, preventing him from realising φ , then it is not the case that he could have realised φ , i.e., $\neg(\delta_1; \dots; \delta_n) \neg vio_G$ attains. Nonetheless, we are obviously more interested in the actions which Aristotle calls “mixed”, actions in which the principle is internal, but which are similar enough to compelled actions that we pardon any violation.

Aristotle gives two examples of such mixed actions. In the first, a tyrant threatens the actor's family with harm if he does not do something “shameful”. In the second, sailors are forced to throw cargo overboard in order to prevent the ship from foundering in a storm. In both cases, the actions are pardonable (perhaps even praiseworthy in the latter) since a greater harm was averted.

In the situation of the ship, for instance, the sailors face a choice between jettisoning the cargo or doing nothing. The safe passage of the cargo is their responsibility, and so they would be accountable for throwing the cargo overboard. But if they do not jettison the cargo, then all will be lost (including the cargo!). We pardon their decision to jettison the cargo, since the alternative is worse. In other words, while they are accountable, they are not blamed.

Of course, to compare the degree of harm (roughly, how bad each violation may be), we would have to complicate the model somewhat, providing a means of comparing violations in distinct worlds so that we may say whether jettisoning the cargo avoids a worse violation than not doing so. In essence, this amounts to adding defeasibility to our model, so that inferences of blame may be “blocked” by defeating obligations, as surveyed in [31]. Obviously, this requires significant work, which is beyond the scope of the present paper. But even without this technical fix, we can get a sense for how mixed actions may be pardonable.

Consider once more the light bulb and switch example, but suppose that Alice has signed an unfortunate non-disclosure agreement (NDA) regarding the state of

the light bulb. Alice and Betty live in a funny old world. In order to reflect this commitment, we must amend the model to satisfy¹⁰

$$M, w_2 \models \text{vio}_a \text{ and } M, w_3 \models \text{vio}_a,$$

since the only way to end up in these two worlds is if Alice breaks the NDA and tells Betty the state of the light bulb. In other words,

$$M, w_0 \models R_a \neg H_G p.$$

Because it is also the case that

$$M, w_0 \models O_G H_G p,$$

it is clear that Alice is between a rock and a hard place. If Alice does nothing (i.e., chooses $\delta_0|_a$), then she is blameworthy for failing to realise $H_G p$. On the other hand, if she chooses to *tell* in the first step, she will be blameworthy for *realising* $H_G p$. It is reasonable to ask whether we should blame her for either or both of these potential outcomes, and it is reasonable to think that our answer depends on the degree of violation. If G 's responsibility for p takes precedence, then we may forgive Alice for violating her non-disclosure agreement, while if the agreement is more important, we may forgive Alice for her role in failing to achieve p —while still blaming G for this failure. In this case, while neither Alice nor Betty is blameworthy for realising $\neg p$, the group $\{a, b\}$ is nonetheless blameworthy.

The situation is even more complicated in the case that a violation can be rectified. Suppose that Alice's non-disclosure agreement has a rectification clause: if she violates the agreement, she can discharge the violation by making amends (perhaps paying a penalty). The situation is shown in Fig. 8. Worlds w_2 and w_3 represent states in which vio_a is true, since in these worlds, Alice has violated her agreement. However, once in these states, she can perform the action *rectify* in order to discharge the violation. Thus, by performing the sequence *tell*; *rectify*, Alice will neither be accountable for the group's responsibility that p , nor for her own non-disclosure agreement, although she *does* enter a violation world (w_2 or w_3) temporarily. The question is whether "all's well that ends well." Should the temporary violation be ignored, even if (for instance) the non-disclosure agreement is more important than the group's obligation to turn on the light?

It is clear that Aristotle's "mixed" actions require some care to represent properly in this formal setting, including the addition of a defeasibility condition to the models, as in [31]. In that case, the violation may be pardonable, even if it is not strictly speaking "forced". The more complicated situation, when a rectifiable but greater violation is temporarily incurred for the sake of a lesser responsibility, requires further consideration.

¹⁰In terms of van de Poel's distinction between "consequentialist" and "deontic" responsibilities [39], it may be more natural to model non-disclosure as a duty not to *tell*, rather than a responsibility not to bring about a certain result. In either case, if Alice brings about $H_G p$ by violating her duty to not execute *tell*, she is accountable for that result, so let us ignore the distinction for now and model the non-disclosure agreement as a duty not to realise $H_G p$.

Fig. 8 In this variation on the light bulb and switch scenario, Alice has a non-disclosure agreement with a rectification clause. If she tells Betty the state of the light bulb, she will violate the agreement, but if she chooses *rectify*, she will realise a non-violation state. Consequently, worlds w_0 , w_1 , w_2 and w_4 are vio_G states, while worlds w_2 and w_3 are vio_a states

7 On Ibo Van De Poel's Four Propositions

Ibo van de Poel [39] has suggested four relations between backward-looking responsibility and forward-looking responsibility.

- P-1 If agent i had a forward-looking responsibility-as-obligation for φ and φ did not occur and $\neg\varphi$ is not caused by exceptional circumstances then agent i is accountable for $\neg\varphi$.
- P-2 If agent i is accountable for φ and has no appropriate excuse why φ is the case then agent i is blameworthy for φ .
- P-3 If an agent i is blameworthy for φ and φ is the case, then agent i is also accountable for φ .
- P-4 If agent i is not accountable for $\neg\varphi$ while φ is not the case and $\neg\varphi$ is not caused by exceptional circumstances, then agent i had no forward-looking responsibility-as-moral-obligation for φ .

We should first note that (P-4) can be restated thus:

If i had a forward-looking responsibility-as-moral-obligation for φ and φ is not the case and $\neg\varphi$ is not caused by exceptional circumstances, then i is accountable for $\neg\varphi$.

Thus, (P-4) is equivalent to (P-1) and so we will omit (P-4) from our considerations.

The first issue in taking up (P-1) is to interpret the phrase “exceptional circumstances”. For our purposes, *non-uniformity* (Section 3) is an exceptional circumstance. That is, if $\delta_1|_G; \dots; \delta_{n+1}|_G$ is *not* uniform at world w , it is possible to “cause” $\neg\varphi$ by executing $\delta_1|_G; \dots; \delta_{n+1}|_G$ just by reaching a world in k steps where $\delta_{k+1}|_G$ cannot be executed. Consider again Fig. 3, in which the sequence $\delta_1|_G; \delta_2|_G$ is non-uniform at w_0 . By executing this sequence, one may end up at world w_3 after one step. At w_3 , the action $\delta_2|_G$ “causes” $\neg\varphi$ —for *any* formula φ —in the sense that

$$M, w_3 \models [\delta_2|_G]\neg\varphi.$$

We regard this way of realising $\neg\varphi$ as extraordinary.

With this definition at hand, we can now show that our definitions of accountability and blameworthiness indeed satisfy (P-1)–(P-3) from [39]. In our terminology, (P-1) can be represented as

$$\models (R_G^n \varphi \wedge E_{\delta_1|G'; \dots; \delta_n|G'} \neg\varphi \wedge U_{\delta_1|G'; \dots; \delta_n|G'} \varphi) \rightarrow \mathcal{A}_{\delta_1|G'; \dots; \delta_n|G'} \neg\varphi. \quad (4)$$

That is, if at world w , G is responsible for φ in n steps and $G' \subseteq G$ causes $\neg\varphi$ by executing $\delta_1|G'; \dots; \delta_n|G' \varphi$ and $\delta_1|G'; \dots; \delta_n|G' \varphi$ is uniform at w , then G' is accountable for $\neg\varphi$.

Note that, by Lemma 2(3), formula (4) is equivalent to

$$\models (R_G^n \varphi \wedge E_{\delta_1|G'} \dots E_{\delta_n|G'} \neg\varphi) \rightarrow \mathcal{A}_{\delta_1|G'; \dots; \delta_n|G'} \neg\varphi. \quad (5)$$

By definition,

$$\mathcal{A}_{\delta_1|G'; \dots; \delta_n|G'}^E \neg\varphi \leftrightarrow E_{\delta_1|G'} \dots E_{\delta_n|G'} \neg\varphi$$

and since

$$(R_G^n \varphi \wedge \mathcal{A}_{\delta_1|G'; \dots; \delta_n|G'}^E \neg\varphi) \rightarrow \mathcal{A}_{\delta_1|G'; \dots; \delta_n|G'} \neg\varphi,$$

formula (5) follows.

Note that uniformity is essential here. Consider again the graph from Figure 3 and assume that w_2 is *not* a violation state for G . The reader can confirm that $M, w_0 \models R_G^2 \top$ and also that $M, w_0 \models E_{\delta_1|G; \delta_2|G} \top$. However, because $M, w_0 \not\models E_{\delta_1|G} E_{\delta_2|G} \top$, it follows that

$$M, w_0 \not\models \mathcal{A}_{\delta_1|G; \delta_2|G} \top.$$

Let us turn our attention to (P-2). We have explicitly interpreted the phrase “ G' has an excuse for φ being the case” in terms of ignorance (lack of distributed knowledge that G' would be accountable for φ) and compulsion (lack of ability to avoid vio_G). Accordingly, we may say that G' has an excuse for φ in case the following proposition holds.

$$\neg K_{G'} \mathcal{A}_{\delta_1|G'; \dots; \delta_n|G'} \varphi \vee \forall_{\delta'_1|G'; \dots; \delta'_n|G'} [\delta'_1|G'; \dots; \delta'_n|G'] \text{vio}_G.$$

Thus, if G' is accountable for φ and has no excuse, then it follows that

$$\mathcal{A}_{\delta_1|G'; \dots; \delta_n|G'} \varphi \wedge K_{G'} \mathcal{A}_{\delta_1|G'; \dots; \delta_n|G'} \varphi \wedge \exists_{\delta'_1|G'; \dots; \delta'_n|G'} \langle \delta'_1|G'; \dots; \delta'_n|G' \rangle \neg \text{vio}_G.$$

It is easy to see that $\mathcal{B}_{\delta'_1|G'; \dots; \delta'_n|G'} \varphi$ follows from this.¹¹

Note that there are instances in which one has an excuse in the above sense, but is nonetheless blameworthy. This arises in our light bulb example (Example 4). Consider the situation in which the group $G = \{a, b\}$ executes the sequence $\delta_2; \delta_0$, as discussed in Example 4. Betty is blameworthy for failing to bring about p by executing this sequence, even though in world w_0 , Betty does not know that she is accountable for $\delta_2; \delta_0$.

Let us turn our attention to (P-3), where the situation is a bit subtler than one may prefer. This proposition requires that blameworthiness entails accountability, but in fact is provable in our system only in the special case that $G' = G$, i.e., one can prove

$$\models \mathcal{B}_{\delta_1|G; \dots; \delta_n|G} \varphi \rightarrow \mathcal{A}_{\delta_1|G; \dots; \delta_n|G} \varphi,$$

¹¹The careful reader will notice that the presumption of $\mathcal{A}_{\delta_1|G'; \dots; \delta_n|G'} \varphi$ is not necessary to infer $\mathcal{B}_{\delta'_1|G'; \dots; \delta'_n|G'} \varphi$ here, since $\mathcal{A}_{\delta_1|G'; \dots; \delta_n|G'} \varphi$ can be inferred from $K_{G'} \mathcal{A}_{\delta_1|G'; \dots; \delta_n|G'} \varphi$.

but the general statement involving an arbitrary subgroup $G' \subseteq G$ does not hold. Let us first prove the theorem in the special case, and then discuss the failure of the general case.

We proceed by induction. The base case is straightforward, since $K_G \mathcal{A}_{\delta|_G} \varphi \rightarrow \mathcal{A}_{\delta|_G} \varphi$ (by the axiom scheme (T')). For the inductive step, note that $\mathcal{B}_{\delta_1|_G; \dots; \delta_{n+1}|_G} \varphi$ entails at least one of the following:

$$R_G^{n+1} \varphi \wedge E_{\delta_1|_G} \mathcal{B}_{\delta_2|_G; \dots; \delta_{n+1}|_G} \varphi, \quad (6)$$

$$K_G \mathcal{A}_{\delta_1|_G; \dots; \delta_{n+1}|_G} \varphi \wedge \exists \delta'_1|_G; \dots; \delta'_{n+1}|_G \langle \delta'_1|_G; \dots; \delta'_{n+1}|_G \rangle \neg \text{vio}_G. \quad (7)$$

In the case that (7) holds, apply axiom scheme (T') again. If (6) attains, apply the inductive hypothesis to conclude

$$R_G^{n+1} \varphi \wedge E_{\delta_1|_G} \mathcal{A}_{\delta_2|_G; \dots; \delta_{n+1}|_G} \varphi.$$

From the definition of $\mathcal{A}$, it follows that either

$$E_{\delta_1|_G} \mathcal{A}_{\delta_2|_G; \dots; \delta_{n+1}|_G}^E \varphi \quad \text{or} \quad E_{\delta_1|_G} (\mathcal{A}_{\delta_2|_G; \dots; \delta_{n+1}|_G}^H \varphi \wedge E_{\delta_2|_G; \dots; \delta_{n+1}|_G} \varphi).$$

In the former case, $\mathcal{A}_{\delta_1|_G; \dots; \delta_{n+1}|_G}^E \varphi$ and hence $\mathcal{A}_{\delta_1|_G; \dots; \delta_{n+1}|_G} \varphi$ immediately follows. In the latter case, we note that

$$\begin{aligned} E_{\delta_1|_G} \mathcal{A}_{\delta_2|_G; \dots; \delta_{n+1}|_G}^H \varphi &\rightarrow \mathcal{A}_{\delta_1|_G; \dots; \delta_{n+1}|_G}^H \varphi \quad \text{and} \\ E_{\delta_1|_G} E_{\delta_2|_G; \dots; \delta_{n+1}|_G} \varphi &\rightarrow E_{\delta_1|_G; \dots; \delta_{n+1}|_G} \varphi \end{aligned}$$

and again $\mathcal{A}_{\delta_1|_G; \dots; \delta_{n+1}|_G} \varphi$ follows.

The problem in the general case $G' \subseteq G$ is that one must show that

$$E_{\delta_1|_G} \mathcal{A}_{\delta_2|_{G'}; \dots; \delta_{n+1}|_{G'}} \varphi \rightarrow \mathcal{A}_{\delta_1|_{G'}; \dots; \delta_{n+1}|_{G'}} \varphi,$$

where the E operator depends on the group G and the accountability relation depends on the subgroup G' . In essence, the incompatibility arises because of the clause (3) discussed in Example 4: if, *at any point* in the execution of $\delta_1|_{G'}; \dots; \delta_n|_{G'}$, the subgroup G' becomes accountable without excuse, then they are blameworthy for their sequence of actions, even though possibly accountable only for the tail of $\delta_1|_{G'}; \dots; \delta_n|_{G'}$. That said, it is still the case that there is a connection between blameworthiness and accountability, even for subgroups. If

$$M, w \models \mathcal{B}_{\delta_1|_{G'}; \dots; \delta_n|_{G'}} \varphi,$$

then there is an $0 \leq i \leq n$ such that

$$M, w \models E_{\delta_1|_G} \dots E_{\delta_i|_G} \mathcal{A}_{\delta_{i+1}|_{G'}; \dots; \delta_n|_{G'}} \varphi.$$

We regard this as quite intuitive. Suppose one recognizes their accountability in the middle of executing $\delta_1|_{G'}; \dots; \delta_n|_{G'}$ and *continues the same execution nonetheless*. Then they are blameworthy for $\delta_1|_{G'}; \dots; \delta_n|_{G'}$. Thus, we regard it as a feature of our formalisation that one may be blameworthy for the whole sequence $\delta_1|_{G'}; \dots; \delta_n|_{G'}$, but accountable only for the tail.

8 Concluding Remarks

In this paper, we have presented a logic that extends PDL by epistemic formulae of the form $\mathcal{K}_i\varphi$, expressing that agent i knows that φ (similarly to [20, 23]), and also by enacted actions, i.e., formulae of the form $[\delta|_G]\varphi$, expressing that φ holds after the execution of δ by group G (similarly to [22, 26, 35, 41]). It turns out that in our logic formulae of the form $E_{\delta|_G}\varphi$, expressing that G has the ability to ensure φ , can be defined as simple abbreviations. Therefore, we can express operators with the same properties as the ones found in other logics of agency, such as coalition logic [33, 34], STIT [22] and ATL [3], but using a simpler semantics. As Ågotnes and Alechina [2] said, PDL has as advantage that it is theoretically well understood, with a range of mathematical and computational tools available without the need for exotic new formalisms. Additionally, as discussed in Section 5.1, the property of Independent Choice (a feature of CEGMs which are used to provide semantics for ATL, Coalition Logic and STIT) is inappropriate for our notion of accountability while CEDL is an appropriate formal setting for capturing our intuitions. Finally, our logic enables us to give a solution to the problem of uniform strategies. With this tool in hand, we propose a formalisation of one notion of forward-looking responsibility, two notions of backward-looking responsibility, and also the relation between them.

We show that, by reasoning about obligations alone, agents are not able to decide to perform the actions that prevent them of reaching a violation state, even in cases where it is possible to do so. The usage of the concepts of forward-looking and backward-looking responsibility in the specification of multiagent normative systems is more appropriate. When agents reason using such concepts, they are able to undertake a course of actions that is more likely to be successful.

Our formalisation of forward-looking and backward-looking responsibility is based on basic “ingredients”: agents’ actions, abilities, obligations and knowledge. Earlier works on the formalisation of responsibility often do not deal with all of these ingredients. For instance, [36] deal only with obligations and agents’ abilities. As we do here, they propose that responsibility should be paraphrased by ‘obligation to ensure’. Their formalisation is done by using a logic wherein one can write formulae of the form $OE_i\varphi$, which stand for ‘it is obligatory that i ensures φ ’. The most interesting feature of this approach is the validity of the scheme $E_iE_j\varphi \rightarrow E_i\varphi$. It expresses that ‘if agent i ensures that agent j ensures φ then i ensures φ ’. This is a useful feature for modelling indirect agency that is not present in our framework. However, Santos and Carmo’s logic is not appropriate to address our problem for two reasons: it does not permit to express agents’ incomplete knowledge about the situation and also does not have actions in its object language. The fact that Betty has incomplete knowledge is crucial in the example presented in Section 2.2, as well as the actions, since the problem is precisely that she needs to decide which action to execute.

With the logic for agent organisation (LAO) [15] Dignum and Dignum propose to formalise responsibility in terms of agents abilities. It seems to be a better alternative than Santos and Carmo’s approach, because the former avoids considering that the agent is backward-looking responsible for a failure in the case the agent is not able

to avoid this failure. However, as in Santos and Carmo's approach, LAO does not have actions in its object language, which means that, again, we cannot address our example using this logic either.

The formalism proposed by Grossi et al. [20] is based on a combination of dynamic and epistemic logics. Therefore, it has actions in the object language and also permits to express incomplete knowledge scenarios. However, this logic does not permit to express agents' abilities. As mentioned above, abilities are important in the definition of responsibility, and it is also important in the definition of obligations.

Possible future works include a couple of questions and improvements. For instance, we did not address decidability, complexity and expressivity of our logic. In particular, a deep comparison analysis between our logic and other logics of agency, such as coalition logic, ATL, STIT [6], etc., has not been addressed here. Another possible extension of our logic that may be promising is the addition of temporal operators. When reasoning about responsibilities, one may want to express statements such as 'Betty must turn on the light before 7 p.m.' This statement means that Betty may fulfill her task by turning on the light before the specified deadline. Such statements cannot be expressed in our language.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

1. Ågotnes, T. (2006). Action and knowledge in alternating-time temporal logic. *Synthese*, 149(2), 375–407.
2. Ågotnes, T., & Alechina, N. (2011). Reasoning about joint action and coalitional ability in k_n with intersection. In *Proceedings of the 12th International Conference on Computational Logic in Multi-agent Systems, CLIMA'11* (pp. 139–156). Berlin: Springer. ISBN 978-3-642-22358-7. <http://dl.acm.org/citation.cfm?id=2044543.2044557>.
3. Alur, R., Henzinger, T.A., Kupferman, O. (2002). Alternating-time temporal logic. *Journal of the ACM*, 5(49), 672–713.
4. Anderson, A.R., & Moore, O.K. (1957). The formal analysis of normative concepts. *The American Sociological Review*, 22, 9–17.
5. Aristotle (2000) *Nicomachean Ethics*. Hackett Publishing Company.
6. Belnap, N., Perloff, M., Xu, M. (2001). *Facing the future*. Oxford University Press.
7. Boudou, J., & Lorini, E. (2018). Concurrent game structures for temporal STIT logic. In André, E., Koenig, S., Dastani, M., Sukthankar, G. (Eds.) *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS 2018* (pp. 381–389). Stockholm. International Foundation for Autonomous Agents and Multiagent Systems Richland, SC, USA / ACM. <http://dl.acm.org/citation.cfm?id=3237443>.
8. Bovens, M. (1998). *The quest for responsibility. Accountability and citizenship in complex organisations*. Cambridge University Press.
9. Bulling, N., & Jamroga, W.J. (2014). Comparing variants of strategic ability: How uncertainty and memory influence general properties of games. *Autonomous Agents and Multi-Agent Systems*, 28(3), 474–518. ISSN 1387-2532. <https://doi.org/10.1007/s10458-013-9231-3>.
10. Chockler, H., & Halpern, J.Y. (2004). Responsibility and blame: a structural-model approach. *Journal of Artificial Intelligence Research*, 22, 93–115.

11. d'Altan, P., Meyer, J.-J.C.h., Wieringa, R. (1996). An integrated framework for ought-to-be and ought-to-do constraints. *Artificial Intelligence and Law*, 4, 77–111.
12. de Lima, T., & Royakkers, L.M.M. (2015). A formalisation of moral responsibility and the problem of many hands. In van de Poel, I.R., Royakkers, L.M.M., Zwart, S.D. (Eds.) *Moral responsibility and the Problem of Many Hands* (pp. 93–130): Routledge.
13. de Lima, T., Royakkers, L.M.M., Dignum, F. (2010). A logic for reasoning about responsibility. *Logic Journal of the IGPL*, 18(1), 99–117. <https://doi.org/10.1093/jigpal/jzp073>.
14. Decker, K., & Lesser, V. (1995). Designing a family of coordination algorithms. In Lesser, V. (Ed.) *Proceedings of the First International Conference on Multi-Agent Systems (ICMAS-95)* (pp. 73–80): The MIT Press.
15. Dignum, V., & Dignum, F. (2007). A logic for agent organizations. In *Proceedings of the 3rd Workshop on Formal Approaches to Multi-Agent Systems (FAMAS)*.
16. Duijf, H., & Broersen, J. (July 2016). Representing Strategies. *ArXiv e-prints*.
17. Fagin, R., Halpern, J., Moses, Y., Vardi, M. (1995). *Reasoning about Knowledge*. The MIT Press.
18. Goldman, A.I. (2009). Social epistemology: Theory and applications. *Royal Institute of Philosophy Supplement*, (64), 1–18.
19. Goodin, R.E. (1986). Responsibilities. *The Philosophical Quarterly*, 42(36), 50–56.
20. Grossi, D., Royakkers, L., Dignum, F. (2007). Organizational structure and responsibility: An analysis in a dynamic logic of organized collective agency. *Artificial Intelligence and Law*, 15, 223–249.
21. Harel, D., Kozen, D., Tiuryn, J. (1984). Dynamic logic. In Gabbay, D., & Guenther, F. (Eds.) *Handbook of Philosophical Logic Volume II – Extensions of Classical Logic* (pp. 497–604): D. Reidel Publishing Company.
22. Herzig, A., & Lorini, E. (2010). A dynamic logic of agency i: Stit, capabilities and powers. *Journal of Logic, Language and Information*, 19(1), 89–121.
23. Herzig, A., Lang, J., Longin, D., Polasek, T. (2000). A logic for planning under partial observability. In *Proceedings of AAAI* (pp. 768–773): AAAI Press/The MIT Press.
24. Herzig, A., Lorini, E., Walther, D. (2013). Reasoning about actions meets strategic logics. In *International Workshop on Logic, Rationality and Interaction* (pp. 162–175): Springer.
25. Jamroga, W.J. (2003). Some remarks on alternating temporal epistemic logic. In *Proceedings of formal approaches to multi-agent systems (FAMAS 2003)* (pp. 133–140).
26. Lorini, E. (2010). A dynamic logic of agency ii: Deterministic dla, coalition logic, and game theory. *Journal of Logic, Language, and Information*, 19(3), 327–351.
27. Lorini, E., Longin, D., Mayor, E. (2014). A logical analysis of responsibility attribution: emotions, individuals and collectives. *Journal of Logic and Computation*, 24(6), 1313–1339. <https://doi.org/10.1093/logcom/ext072>.
28. Meyer, J.-J.C.h. (1988). A different approach to deontic logic: Deontic logic viewed as a variant of dynamic logic. *Notre Dame Journal of Formal Logic*, 29(1), 109–136.
29. Meyer, J.-J.C.h., & Wieringa, R. (1993). *Editors. Deontic Logic in Computer Science*. Chichester: Wiley.
30. Miller, S. (2007). Against the collective moral autonomy thesis. *Journal of Social Philosophy*, 38(3), 389–409.
31. Nute, D. (1997). *Defeasible deontic logic*. Kluwer Academic Publishers.
32. Osborne, J., & Rubinstein, A. (1994). *A course in game theory the*. MIT Press.
33. Pauly, M. (2001). *Logic for Social Software*. PhD Thesis, ILLC University of Amsterdam.
34. Pauly, M. (2002). A modal logic for coalitional power in games. *Journal of Logic and Computation*, 12(1), 149–166.
35. Royakkers, L.M.M. (1998). *Extending deontic logics for the formalisation of legal rules*. Kluwer Academic Publishers.
36. Santos, F., & Carmo, J. (1996). Indirect action, influence and responsibility. In Brown, M.A., Carmo, J., van Rijsbergen, C.J. (Eds.) *Deontic Logic, Agency and Normative Systems: Workshops in Computing series*. Springer.
37. Thompson, D.F. (1980). Moral responsibility and public officials. *American Political Science Review*, 74, 905–916.
38. van de Poel, I.R. (2011). The relation between forward-looking and backward-looking responsibility. In Vincent, N., van de Poel, I.R., van den Hoven, J. (Eds.) *Moral responsibility. Beyond Free Will and Determinism* (pp. 37–52): Springer.
39. van de Poel, I.R. (2015). Moral responsibility. In van de Poel, I.R., Royakkers, L.M.M., Zwart, S.D. (Eds.) *Moral responsibility and the Problem of Many Hands* (pp. 12–49). Routledge.

40. van Fraassen, B.C. (1973). The logic of conditional obligation. In Bunge, M. (Ed.) *Exact Philosophy* (pp. 151–172). D. Reidel Publishing Company.
41. Wieringa, R.J., & Meyer, J.-J.C.h. (1993). Actors, actions, and initiative in normative system specification. *Annals of Mathematics and Artificial Intelligence*, 7, 289–346.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.